COMPUTATIONAL METHODS FOR MEASUREMENT OF VISUAL ATTENTION FROM VIDEOS TOWARDS LARGE-SCALE BEHAVIORAL ANALYSIS

A Dissertation
Presented to
The Academic Faculty

By

Eun Ji Chong

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing, College of Computing

Georgia Institute of Technology

May 2020

COMPUTATIONAL METHODS FOR MEASUREMENT OF VISUAL ATTENTION FROM VIDEOS TOWARDS LARGE-SCALE BEHAVIORAL ANALYSIS

Approved by:

Dr. James M. Rehg, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Agata Rozga
School of Interactive Computing
Georgia Institute of Technology

Dr. Gregory D. Abowd
School of Interactive Computing
Georgia Institute of Technology

Dr. Irfan Essa
School of Interactive Computing
Georgia Institute of Technology

Dr. Yaser Sheikh
Robotics Institute
Carnegie Mellon University

Date Approved: January 9, 2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Visual attention is one of the most important aspects of human social behavior, visual navigation, and interaction with the world, revealing information about their social, cognitive, and affective states. Although monitor-based and wearable eye trackers are widely available, they are not sufficient to support the large-scale collection of naturalistic gaze data in face-to-face social interactions or during interactions with 3D environments. Wearable eye trackers are burdensome to participants and bring issues of calibration, compliance, cost, and battery life. The ability to automatically measure attention from ordinary videos would deliver scalable, dense, and objective measurements to use in practice.

This thesis investigates several computational methods to measure visual attention from videos using computer vision and its use for quantifying visual social cues such as eye contact and joint attention. Specifically, three methods are investigated. First, I present methods for detection of looks to camera in first-person view and its use for eye contact detection. Experimental results show that the presented method can achieve the first human expert-level detection of eye contact. Second, I develop a method for tracking heads in a 3d space for measuring attentional shifts. Lastly, I propose spatiotemporal deep neural networks for detecting time-varying attention targets in video and present its application for the detection of shared attention and joint attention. The method achieves state-of-the-art results on different benchmark datasets on attention measurement as well as the first empirical result on clinically-relevant gaze shift classification.

Presented approaches have the benefit of linking gaze estimation to the broader tasks of action recognition and dynamic visual scene understanding, and bears potential as a useful tool for understanding attention in various contexts such as human social interactions, skill assessments, and human-robot interactions.

CHAPTER 1
INTRODUCTION

Visual attention is a critically-important aspect of human social behavior, visual navigation, and interaction with the 3D environment, and where and what people are paying attention to reveals a lot of information about their social, cognitive, and affective states. While monitor-based and wearable eye trackers are widely-available, they are not sufficient to support the large-scale collection of naturalistic gaze data in contexts such as face-to-face social interactions or object manipulation in 3D environments. Wearable eye trackers are burdensome to participants and bring issues of calibration, compliance, cost, and battery life.

Besides the benefit of scalability, using automated measure of visual attention is also valuable for its ability to provide dense and objective assessment on one's gaze behavior. Unlike manual coding or qualitative questionnaire, automated methods can produce fine-grained signals from video at much lower cost, which are more suitable for capturing time-sensitive information. It is also less susceptible to personal biases provided that the bias present in the training data can be overcome by including labels from diverse set of annotators.

As a result, it can have significant practical impact on various applications across different domains. For example, applications in clinical and social psychology include analysis of people's use of gaze during multi-person interaction when they signal each other or establish rapport. It can also be particularly useful in identifying social behaviors among individuals with developmental disorders such as autism by measuring their social gaze for the purpose of screening, diagnosis and treatment of the disorder. Other applications could be skill assessment in lab experiment, surgery, or job interviews as what people look at is closely related to how well they perform in completing the goal task. In addition, it can

benefit the development of social intelligence for robots, enabling them to interact naturally with people using nonverbal social signals.

There exist a spectrum of technologies designed to measure visual attention. Monitor-based eye tracker is one of the most widely and commercially available options and is convenient to use for controlled lab experiments as the experimenter can fully control what is displayed to the viewer. Wearable eye tracker may be used to allow for free movement and interaction with real world in certain settings. However, if we could avoid any type of such devices and measure gaze behavior using ordinary recorders such as camcorder or mobile phone, that would be the most scalable, ecologically valid and therefore best suited option for large-scale deployment in the study and analysis of natural human behavior. The biggest challenge in this scenario is the drop in precision caused by the lack of dedicated imager for the eye balls. This thesis aims to bridge this gap by developing specialized computer vision methods and datasets for non-invasive attention estimation.

1.1 Objective

In this thesis, I investigate different computational methods for measuring real-world human visual attention in video to support large-scale behavioral analysis and demonstrate its use for quantifying meaningful social behaviors such as eye contact and joint attention.

First, I start with a simple setting in which a point-of-view camera worn by a social partner is used to detect bids for eye contact made by the subject. In this setting, detector is making a binary classification of the person's looks to the camera and offered an advantage of better handling occlusion and ambiguity during dynamic face-to-face interactions. Then, I include an additional viewpoint which is a tripod-mounted camcorder to capture the holistic view of the scene and develop a method for tracking 3D head pose by fusing measurements from the two disparate views. With this approach, I extend beyond eye contact and enable the detection of attentional shifts from a known 3D object to social partner's face under the assumption that the head motion follows the direction of focus of

attention. Lastly, I consider the most general case of attention measurement where we make no assumptions regarding the target object and support any types of viewpoint. Presented method has the goal of predicting the location of visually attended region in every video frame, given a track of an individual's head, including the cases where such target is out of frame, in which case the model should correctly infer its absence.

It should be noted that it is inevitable for the non-invasive methods to have a lower accuracy as compared to that of wearable eye trackers due to the ambiguities caused by low resolution and occlusions. With these challenges in mind, I aim to achieve the performance equivalent or close to human-level reliability because the automated methods would then be ultimately good enough to use in place of manual coding.

1.2 Thesis Statement

Visual attention during face-to-face interactions is non-invasively measurable from videos to support large-scale data analysis.

1.3 Overview

My research is organized into three main topics: Detection of Eye Contact in Egocentric View (Chapter 2), Head Pose-Based Attention Shift Measurement (Chapter 3), Generalized Attention Target Detection (Chapter 4). I present an overview of each topic in this section.

1.3.1 Detection of Eye Contact in Egocentric View

This chapter presents novel approaches to eye contact detection in face-to-face social interactions in which the social partner wears a point-of-view camera to capture eye contact bids. By analyzing facial regions and modeling the appearance change upon head motion, the method can accurately identify the onset of the subject's looks to their social partner's eyes. I introduce deep learning method that can detect looks to the eyes while implicitly learning the condition of eye contact under various angles. I present an automated

system for eye contact detection that solves the sub-problems of pose estimation and end-to-end feature learning using deep convolutional neural networks. The trained model is extensively evaluated using real-world datasets in clinical assessment setting and yields performance equivalent to expert human raters.

1.3.2 Head Pose-Based Attention Shift Detection

This chapter describes an approach of continuously capturing and tracking 3D head pose in a tabletop social interaction between two people. The approach utilizes a fixed room camera in conjunction with a dynamic camera worn on social partner which are arranged to simultaneously record the subject's face along with a known 3D object. The proposed system performs head tracking and pose estimation under a simplified multiple-hypothesis tracking framework and provides 3D states of each entity in a unified coordinate system every time step. Then, I present how an attention shift behavior can be inferred from the obtained measurements.

1.3.3 Generalized Attention Target Detection

This chapter addresses the problem of detecting attention targets in video by proposing a novel spatiotemporal deep learning architecture that models the dynamic interaction between the scene and head features. The method takes a third-person video and identifies where each person present in each frame is looking, and correctly handle the out-of-frame case. A novel dataset containing video sequences of annotated dynamic gaze tracks of people in diverse situations is created to make the model learn the temporal change of attention. Furthermore, I demonstrate the value of this approach by using the predicted attention map from the model for social gaze recognition tasks. Experiments show that the method achieves state-of-the-art results on multiple benchmark datasets and a novel social gaze recognition task.

My dissertation makes the following contributions:

- I presented a method for detecting eye contact from wearable camera worn by a social partner and demonstrated that the method can achieve accuracy equivalent to that of trained human coders
- I presented a method for tracking heads in a 3D space during dynamic social interactions between two people, and showed that it can be used for detecting attentional shifts
- I presented a method for detecting attention targets in a 3rd-person video and demonstrated its utility for different social gaze behavior recognition tasks
- I contributed to the creation of a novel dataset that consists of complex and dynamic patterns of real-world gaze behavior
- I contributed to the collection and analysis of a large-scale dataset of naturalistic child-adult social interactions on which the presented methods are applied and evaluated

CHAPTER 2

DETECTION OF EYE CONTACT IN EGOCENTRIC VIEW

2.1 Introduction

Gaze behavior is a key foundation of face-to-face social interaction. Eye contact, the act of looking another person in the eyes, is one of the earliest social skills to emerge in development [15, 16], and studies have shown that infants are tuned to looking at faces from birth [17, 18]. Eye contact serves multiple important functions in social communication, including the establishment and recognition of relationships between partners and the expression of interest and attentiveness [19, 20]. Moreover, it is a core component of joint attention, in coordination with other gestures [21]. Atypical use of eye contact and abnormal gaze patterns are often part of a list of red flags for numerous medical and/or psychiatric conditions, including Autism Spectrum Disorder (ASD) [22], Fragile X syndrome [23], ADHD [24], Williams Syndrome [25], social anxiety/behavioral inhibition [26], and oppositional defiant disorder [27]. In particular, decreased eye contact is one of the formal diagnostic criteria for ASD [28, 13], and is also highlighted during early screening and treatment.

As a result of the critical importance of gaze, a variety of technologies have been developed to automate the measurement of gaze behavior, of which eye tracking is the best known example. Conventional monitor-based eye tracking is unsuitable for measuring the contingent real-world aspects of social gaze during face-to-face interactions. While wearable eye trackers can be utilized to measure gaze behavior in adults [29, 30, 31] and infants [32, 33], they are both expensive and burdensome to the subject. The need to wear and calibrate eye tracking hardware can be a tremendous challenge to subjects with compliance, distraction or fatigue issues, and this can affect both the yield and quality of the data. Infants, young children, and individuals with health problems are examples of subject groups

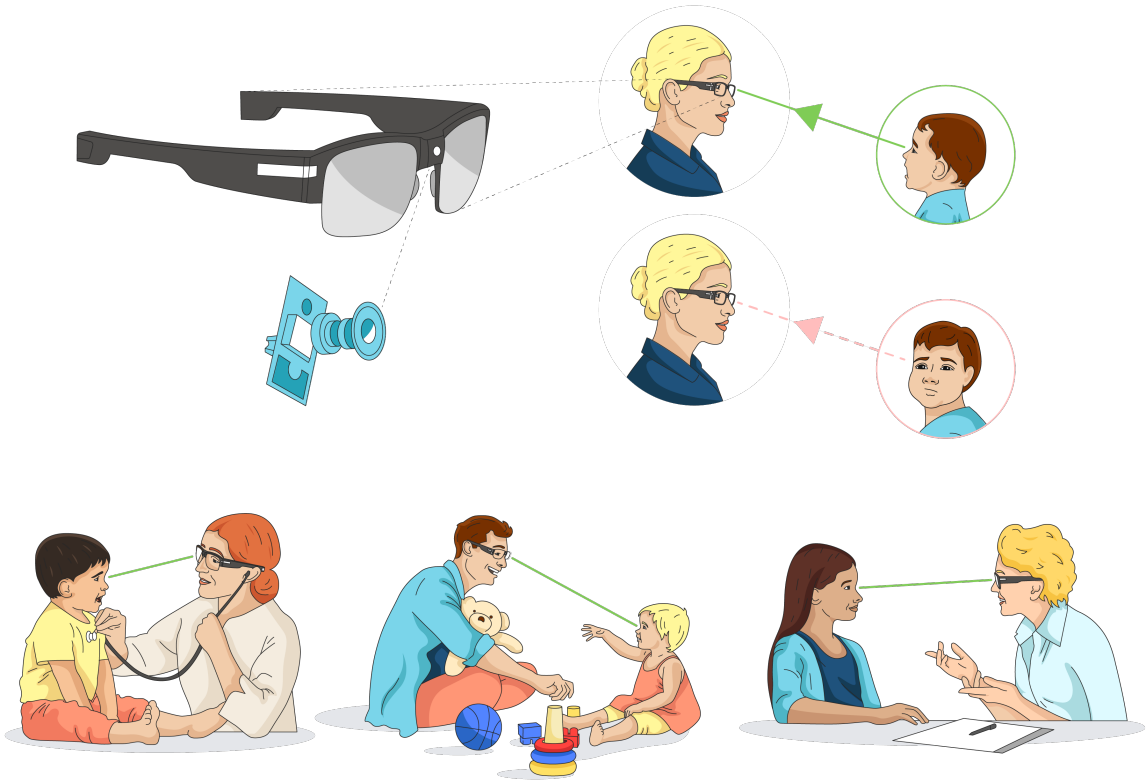


Figure 2.1: **Overview of the approach.** Wearable glasses with a small outward-facing camera embedded in the bridge are used to record the face of the camera wearer's social partner. By virtue of its placement, gaze during eye contact is directed towards the camera, and is captured in video, enabling automated detection. Due to its ease of use, the approach can be widely deployed in a variety of settings, as illustrated in the figure, for which eye tracking may be infeasible due to cost, burden, compliance, or distraction issues.

that are likely to have such difficulties. Moreover, since the eye tracker only provides the point of gaze in a captured video recording, manual annotation must be performed in order to identify the gaze targets, limiting the scalability of the approach.

We have pioneered a novel, scalable, low-burden approach to automatically detecting moments of eye contact between individuals during face-to-face interactions [34, 35], illustrated in Fig. 4.4. The examiner wears a low-cost pair of glasses with a point-of-view (PoV) camera embedded in the bridge, which serves as a video recorder. By virtue of its placement, the subject will be looking directly towards the camera any time they are making eye contact with the examiner, facilitating automatic detection using computer vision methods. In our approach, the subject is completely unencumbered and the examiner burden is low

since the glasses are light-weight and unobtrusive.

While human raters can achieve levels of agreement above 90 percent when gaze coding PoV videos [36, 37], the accuracy achieved in prior works [34, 35, 38, 39] is well below this level of performance, making automatic coding unusable by researchers and practitioners as a measurement tool. This paper addresses this challenge by exploring three directions. First, we hypothesize that modern deep learning architectures can effectively exploit a large dataset of 4.7M human-annnotated eye contact events in PoV video. However, while our dataset is large by any standard, it contains only around 100 unique subjects. In contrast, datasets for face detection, recognition, and other tasks, which have been shown to yield high performance when using deep models [40, 41, 42], contain orders of magnitude more variability. We hypothesize that we can close this gap by using task transfer learning from additional datasets that model the relationship between head pose and eye gaze direction, which is central to our task. Third, we hypothesize that the frequency and duration of moments of eye contact identified by our automated method will correlate with measures of social impairment among subjects with ASD. Establishing this hypothesis validates the feasibility of fully-automated eye contact coding using our approach.

2.2 Related Work

There are three categories of relevant prior work. First, we compare to the small number of previous works that addressed eye contact detection. Second, we describe recent works on large scale eye tracking that also make use of deep models for appearance analysis of eye regions. Third, we briefly review the use of classical eye tracking methods in autism.

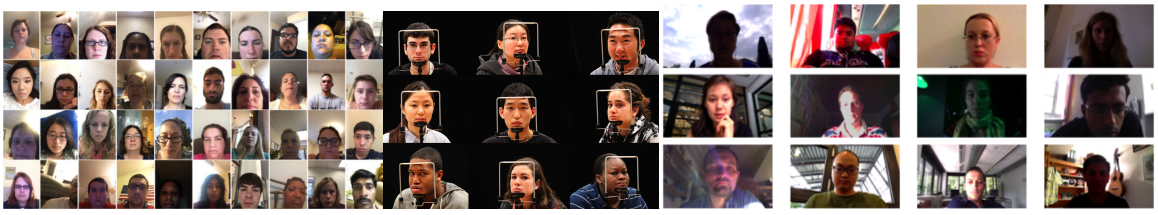
Eye Contact Detection: A few studies previously addressed the direct estimation eye contact from video and constitute the closest related work [38, 43, 44, 45]. The most relevant are our previous papers [43] which introduced the POV camera paradigm for eye contact detection and the work of [38] that introduced a novel detection approach based on deep learning which dramatically improves the detection performance of [43]. It also pre-

sented the first experimental results for children with ASD, demonstrating that diagnostic status does not have a significant impact on detection performance. However, this work did not present its competence against human coders.

Shell et al. [44] developed an approach to eye contact detection based upon classical gaze tracking methods, using IR diodes on a pair of glasses to create glints on the eyes of the social partner. The reliance on special IR illumination greatly limits the usefulness of the method for naturalistic interaction. In their work on gaze locking [45], Smith et al. addressed a different application of eye contact, namely the use of gaze to an embedded camera as a user-interface technology in an internet-of-things context. Their approach predated the wide-spread use of deep models and their dataset consisted of subjects in a chin-rest, with the consequence that they could not present results for naturalistic interactions.

Appearance-Based Gaze Estimation and Existing Datasets: Traditional approaches to gaze estimation utilize active IR illumination to both create glints on the surface of the eye and reliably segment the pupil opening using a variety of dark and light pupil methods [46]. Recently, a number of investigators have explored alternative approaches to gaze estimation using appearance-based methods that analyze the eye region in conventional RGB images and avoid the use of structured illumination [47, 48, 49, 50]. We share with these methods the observation that the analysis of the eye region pattern in combination with head pose is a viable alternative to conventional gaze tracking technology. The major difference is that these prior works address the traditional eye tracking goal of determining the user's point of gaze on a display surface, as motivated by the widespread availability of user-facing cameras in tablets and laptop screens. These methods cannot be applied directly in our context of naturalistic face-to-face social interactions.

There are several publicly-available gaze datasets available to the research community. However, as these datasets were collected from interactions with screens, they are not suitable for building a model that describes face-to-face social gaze behaviors. Example images



(a) GazeCapture [48] (b) Columbia Gaze [45] (c) MPIIGaze [50]

Figure 2.2: **Examples of publicly available gaze datasets.** As these datasets were collected from interactions with screens, they are not suitable for building a model that describes face-to-face social gaze behaviors.

Figure 2.3: **Sample child faces in our dataset.** Green box indicates ground truth eye contact, red indicates no eye contact. Notice the diversity and richness in the children's facial expressions and head pose during natural social interactions.

for the major public datasets are illustrated in Fig 2.2. As these images demonstrate, the facial expressions and poses obtained from adult subjects interacting with screens are much less diverse than the variations in children's appearance and pose that we encounter in our setting (see Fig 2.3 for examples).

Gaze Behavior and Autism: A significant amount of prior work has used eye tracking to investigate differences in patterns of looking in individuals with autism, such as reduced looks to social stimuli [51]. For example, toddlers with ASD spend more time looking at geometric shapes than human biological motion [52], children with ASD devote less attention to faces while watching videos of social interactions [53], and both adults [54] and children [55] with ASD show preferential fixations to the mouth than to the eyes when

viewing social scenes. However, all of these studies have been conducted in a highly-controlled environment in which the subjects were passively viewing a monitor screen for a short period of time. At present, we still lack a comprehensive understanding of similarities and differences in gaze behavior associated with autism [56]. Our work on eye contact can potentially complement this existing literature by providing insight into patterns of looking within a naturalistic social context.

Head-mounted eye tracking systems provide an alternative to monitor-based studies of gaze behavior and have been used in a limited number of studies involving children with ASD [57, 58]. Findings from this work exhibit concordance with monitor-based gaze studies, but have also identified novel gaze patterns [57], suggesting that more research is needed to understand gaze behavior in ecologically valid settings. A basic problem with using any form of eye tracking to analyze face-to-face gaze is the need to identify the gaze target given the estimated gaze direction. In other words, wearable eye tracking gives the location of the point of regard in a POV image, but does not directly answer the question of what gaze target is present at that location. This difficulty substantially increases the complexity of a fully-automated behavior measurement system based on wearable eye tracking. These issues, along with the challenges of compliance in requiring children to wear special hardware [59], have limited the broad-scale applicability of this approach.

We note that a final, classical approach to obtaining social gaze measurements, including measurements of eye contact, is to manually annotate videos recorded in the lab setting or even home videos [60]. While this method is completely non-invasive and naturalistic, it is extremely time consuming and is subject to human error. Our previous work has shown that video from POV cameras is an effective medium for human annotation of children's social gaze [36]. We utilize such annotations to construct the training and testing sets for our experiments.

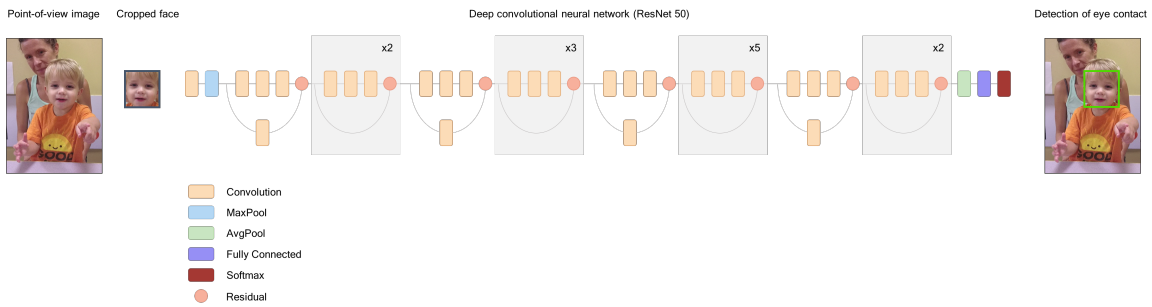


Figure 2.4: **Deep neural network layout.** Given a frame extracted from point-of-view camera, subject's face is automatically detected and cropped as an input to the deep neural networks. ResNet 50 network architecture is used to compute the features from facial image via a series of convolutions. At the end of the network, features are combined through average pooling and fully connected layer, and the softmax operation produces the final eye contact score. Using this score, algorithm can decide if the input face is an eye contact.

2.3 Methods

2.3.1 Dataset

The dataset used in this study was collected at two institutions between 2015 and 2018; Neurotypical subjects were recruited at the Georgia Tech Child Study Lab in Atlanta, GA (GT), and subjects with ASD were recruited through the Center for Autism and the Developing Brain in White Plains, NY (CADB). All caregivers provided written consent and the Georgia Institute of Technology and Weill Cornell Medicine IRB approved the study.

Data collection setup

Each subject completed a set of naturalistic, semi-structured play interactions with a trained examiner. The play interactions, described in more detail below, are designed to elicit nonverbal communication behaviors that are present in typical development, but are often less prevalent in children with autism. Interactions took place at a table with the subject sitting across from the examiner, either on their parent's lap or independently, in order to facilitate data collection. The examiner wore a pair of commercially-available glasses - Pivothead Kudu - that provides continuous high resolution capture of the subject's face. The

lenses were removed from the glasses to provide an unobstructed view of the examiner's eyes. A stationary camcoder mounted on a tripod was positioned to capture a holistic view of the scene, including the table and all subjects.

Play interaction 1: Early Social Communication Scales

The Early Social Communication Scales (ESCS) [12] is a structured interaction in which the examiner presents a series of toys in order to elicit nonverbal social communication behaviors (e.g., use of pointing, reaching or eye contact to initiate joint attention to a toy or to request). Administration of ESCS was slightly modified appropriately for the study as follows. As eye contact, rather than other social communication behaviors, was the main focus of the current study, ESCS administration was modified in order to elicit more instances of direct gaze. For example, for items that require the child to make a bid to the examiner in order to obtain the toy, examiners would give the toy only when the child made eye contact, not in response to other bids such as pointing or verbally asking, with the exception that if the child lost interest in the toy or became agitated prior to making eye contact the examiner would give them the toy without requiring any bid. Administration was also modified to remove materials that are meant to be used close to the examiner's face (i.e., hat, comb and glasses), to prevent ambiguity as to whether the child is looking at the examiner's eyes or at the toy. The ESCS was administered to children 60 months of age and younger.

Play interaction 2: Brief Observation of Social Communication Change

The Brief Observation of Social Communication Change (BOSCC) [14] is a semi-structured play-based interaction between the subject and their interaction partner. Subjects in this study completed a modified version of the BOSCC at the table; The BOSCC consisted of two, four-minute segments of play that were followed by two, two-minute segments of snack or conversation based upon the age of the subject. For the play segments, the subject

was presented with a box of toys and asked to select one toy. If the subject did not select a toy, the examiner selected one for themselves and the subject to play with. The subject was only permitted to have one toy on the table at a time, but the subject could select a different toy at any time. During snack segments, the examiner presented two clear containers with different snacks, and the subject selected which snack they would like. Children were given small portions of snack in order to create opportunities for additional requesting. The BOSCC is designed to measure change over time, and its inclusion in our dataset serves to increase the diversity of the gaze behavior.

Subjects

66 children (55 male) who were suspected of having ASD participated at CADB and 58 typically developing (TD) children (36 male) participated at Georgia Tech, in a study that was designed to validate the feasibility of using PoV glasses to capture gaze behavior in young children (18-60 months, M=36.48 months). Subjects in the young children sample who were suspected of having ASD were evaluated with the Autism Diagnostic Observation Schedule (ADOS-2) [13]. TD subjects were screened for developmental delays with the the CSBS-DP [61] or M-CHAT [62]. Three TD subjects were excluded from all analyses due to technical issues. Three subjects scored in the low range of concern for ASD and one subject did not complete the ADOS-2 and these four subjects were not included in analyses comparing ASD to TD groups. Three TD subjects were excluded from analyses comparing across groups because of concern for developmental delay. Subjects completed the ESCS and the BOSCC in randomized order. A subset of subjects (n=14) completed a follow-up session within a year of their first visit, and another subset (n=27) completed an additional BOSCC with their parent as the examiner. Eight subjects completed a second parent-led BOSCC at their follow-up session. In total, the number of sessions is 167. All subjects in the young children sample, regardless of diagnostic status, were utilized in training and validating the eye contact model. Thus, the dataset consists of 66 ASD and 55

Table 2.1: **Demographics and descriptive statistics.** CSS total calibrated severity score from ADOS, CSS SA ADOS calibrated severity score for social affect, CSS RRB ADOS calibrated severity score for restricted and repetitive behaviors, CBCL Total score for internalizing and externalizing problem behavior. Age, CSS and CBCL scores expressed as Mean (Standard Deviation). Higher scores indicate more significant impairment/problem behavior.

	TD young children sample	ASD young children sample	ASD child and adolescent sample
N	55	66	15
Males	36 (65%)	55 (83%)	12 (80%)
Age (months)	27.45 (5.84)	44.00 (11.11)	95.56 (32.60)
Score			
CSS total	N/A	7.97 (2.19)	7.43 (1.74)
CSS SA	N/A	7.49 (2.30)	7.64 (1.74)
CSS RRB	N/A	8.19 (1.76)	6.79 (2.83)
CBCL Total T score	43.16 (9.02)	58.05 (15.35)	59.00 (8.88)
Race			
White/Caucasian (%)	60	61	60
Black/African American(%)	22	4	0
Asian/Pacific Islander (%)	0	15	7
More than one race (%)	13	14	26
Other/Unknown (%)	5	6	7
Hispanic/Latino ethnicity	4 (7%)	15 (23%)	3 (20%)
Maternal education			
Some high school	1	0	0
High school diploma/GED	3	3	2
Some college	8	4	0
College/technical degree	24	28	5
Graduate school degree	19	30	8
Unknown	0	1	0

TD children (see columns 1 and 2 in Table 2.1 for detailed demographic information).

Video coding of eye contact

Mangold International's INTERACT video annotation software [63] was used by coders to flag frame-level onsets and offsets of eye contact during ESCS and BOSCC protocols. The ESCS sessions from 10 subjects (5 TD, 5 ASD) were annotated by 8 independent raters to establish reliability (mean $\kappa = 0.886$). In addition, 3 raters (1 rater in common with the first group) annotated the BOSCC videos from 8 subjects (4 TD, 4 ASD), in

order to test reliability on a different play protocol (mean $\kappa = 0.903$). The remaining ESCS and BOSCC sessions were annotated by single raters. The video segments from the 10 ESCS and 8 BOSCC sessions which were annotated by multiple raters constitute the *validation set*, which was used to test generalization performance after training. The remaining single-rater sessions from 103 subjects comprise the *training set*. The validation set has no overlap with the training set and is representative of the total sample in terms of age, race, diagnosis, gender and autism severity. Note that we have shown in prior work that human coding of eye contact from PoV video can be achieved with greater reliability than from a standard video recorder [36], thus validating our data annotation approach. This work also demonstrated that wearing the glasses did not impact the frequency of eye contact bids in a sample of 2-to-4-year-olds.

2.3.2 Data preparation

PoV video frames were decoded at 30 frames per second (capture rate) and saved to disk. In each frame, the subject's face was detected and recognized following the procedure from [38]. In the training set, each frame is labeled by a single rater, with 1 for eye contact and 0 otherwise. In the validation set, each frame has annotations from multiple raters and the majority vote is used as the ground truth label. The datasets consist of 4,339,879 frames (281,152 with eye contact) for training and 353,924 frames (25,112 with eye contact) for validation.

2.3.3 Training algorithm

We used a deep convolutional neural network (CNN) with a ResNet 50 backbone architecture [64] as our classifier model (see Fig. 4.3). The inputs consist of cropped face regions, resized to 224 by 224 pixels. We used a two-stage training process to support task transfer learning. In the first stage, training on three public datasets enables the model to learn the relationship between head pose and eye gaze direction. The model is trained to regress

the 3D gaze direction based on MPIIGaze [65] and EYEDIAP [66] datasets and 3D head pose with the SynHead [67] dataset, using an L2 regression loss. Convergence is defined as reaching $< 6^{\circ}$ mean absolute error on gaze angle and head pose. The model is then fine-tuned using the training dataset, in order to learn the condition of eye contact and capture the details of children's facial appearance. Fine-tuning was done across the last two blocks of the ResNet layers, using cross entropy loss with a reweighting factor of 0.1 which is multiplied by the loss of the over-represented class in order to compensate for the class imbalance (eye contact presence vs. absence ratio) in our dataset. Backpropagation with a learning rate of 0.005 under Adam optimization is used to update network weights for every mini-batch of 256 samples until it has seen 3 epochs of training data with augmentation.

Data augmentation during training consists of a combination of horizontal flip, color jitter (brightness, contrast, saturation $\leq \pm 20\%$), blur (guassian kernel size ≤ 0.6), and face bounding box rescale ($\leq +20\%$), each of which takes effect at a probability of 0.5. Augmentation parameters are uniformly sampled within the given range. Note that our prior work [38] used an architecture that incorporated multi-task learning, by forcing the network to predict the head pose and the eye contact label during training. A finding from this work is that the transfer learning approach is more effective.

2.3.4 Evaluation

We performed three experiments to evaluate the performance of our approach. The first experiment evaluates the per frame prediction accuracy of our method. The second experiment evaluates the inter-rater reliability of the deep learning model with respect to a set of human raters. The third experiment tests whether the direct application of our eye contact detector can replicate findings about eye contact behavior in two previously-reported studies which used manual coding. These experiments provide three sources of converging evidence for our primary hypothesis: That automatic detection of eye contact via a deep learning classifier yields performance which is equivalent to the accuracy of human coders,

making automatic video coding viable for research studies in social communication.

Frame-level accuracy

We compute the precision-recall (PR) curve on the validation dataset as a function of the classification threshold, using the majority vote of the human coders as the ground truth.

$$precision = \frac{\text{true positive}}{\text{true positive + false positive}}$$

$$recall = \frac{\text{true positive}}{\text{true positive + false negative}}$$

where 'true positive' is the number of correctly predicted eye contact, 'false positive' is the number of incorrectly predicted non eye contact, 'false negative' is the number of incorrectly predicted eye contact.

As the neural network outputs a score S per frame (taken from the softmax layer), a precision-recall (PR) curve is generated by choosing a fixed threshold score t such that the prediction $\hat{y}$ for each image is defined as $\hat{y} = S \geq t$, then sweeping t in the interval 0-1. Due to the imbalance between the two classes (eye contact vs. other) in our dataset, PR curve is better suited for evaluation than Receiver Operating Characteristic, as PR does not take into account abundant true negatives. Maximum F1 score $= \frac{2 \times P_t \times R_t}{P_t + R_t}$ and Average Precision (AP) $= \sum_t (R_t - R_{t-1}) \times P_t$ are also reported as summary statistics for PR curve (P_t and R_t are the precision and recall, respectively, at threshold t).

Reliability with human raters

For this evaluation, we threshold the model's output and perform post-processing to predict an eye contact label for each frame of an input video sequence. Specifically, we used a decision threshold of 0.9 in order to predict the presence or absence eye contact in each frame based the softmax output of the deep model classifier. In addition, we performed temporal smoothing as a post-processing step in order reduce noise caused by unwanted

events such as face detection failure, motion blur, and eye blinks, we remove outliers and merge short segments through a sliding window scheme. The classifier decision threshold and the window sizes for outlier removal (5) and merging (6) were chosen via grid search on a held-out training sample, by maximizing detection accuracy while minimizing the event-level eye contact count difference between the estimate and human coding.

We treat the learned model as an additional (automated) video coder, and compute the agreement between the algorithm and the human raters on the validation dataset. We use Cohen's kappa score [68] to measure the inter-rater reliability between the model and each of the raters, and between all pairs of raters, where an average kappa score greater than 0.8 is usually considered to be sufficient to establish reliability for a group of raters.

In addition, we use two one-sided tests (TOST) [69] to statistically test the hypothesis of equivalence between the human annotator group and the automatic coding algorithm. In TOST, the null hypothesis is a sample mean difference greater than Δ, and the alternative hypothesis is equivalence between the classes:

$$H_0: \ m_{hd} - m_{hh} < -\Delta \ \text{ or } \ m_{hd} - m_{hh} > \Delta$$
$$H_1: \ -\Delta < m_{hd} - m_{hh} < \Delta$$

where m_{hc} is the mean of kappa scores of all human-detector pairs, m_{hh} is the mean of Cohen's kappa scores between all human pairs, and Δ is the equivalence boundary.

Replication of prior studies

Our final evaluation assesses the impact of replacing human coding with computer coding of eye contact in prior observational studies. We repeat the data analysis for the two studies described in [70, 37], using eye contact statistics obtained from applying our automated method to the original video files, as an alternative to the manual coding originally performed by the authors. For both studies, we compute eye contact frequency (eye contact event counts per minute) and duration rate (eye contact duration divided by the adminis-

tration time) using the algorithm's output, and repeat the statistical tests as in the original papers.

The study in [70] examined differences in eye contact rates between ASD and TD young children during the toy-spectacle tasks of the ESCS in consideration of temporal-contextual factors such as activation and possession of the toy. Samples used in training the eye contact model overlap partially with the samples used in [70]. In order to eliminate any potential bias, we removed the overlapping subjects (N=47) from the training set and retrained the eye contact model on 56 subjects (instead of 103) for use in this experiment.

The study in [37] investigated increased eye contact in individuals with ASD during conversation as compared to play in the BOSCC protocol in an additional group of 15 older children and adolescents with ASD (3 female, 5-13 years, M=8 years). Table 2.1, column 3 provides detailed demographic information. We reproduce the analysis of 'Sample 2' from [37] only, as 'Sample 1' was not available.

2.3.5 Correlation analysis

We performed an additional experiment to examine the correlation between automatically measured eye contact statistics (frequency and duration rate) and symptom severity, using the ASD young children sample. Given the importance of eye contact for overall social communication skills in toddlers [21, 20, 19], it was expected that frequency and duration of eye contact would demonstrate a negative correlation with social impairment among autism subjects.

Symptom severity was assessed using two social affect scores derived from the ADOS and the BOSCC. First, the ADOS Calibrated Severity Scores for Social Affect (ADOS CSS SA) were computed from the ADOS [71] results. In addition, the BOSCC Social Affect (BOSCC SA) scores were coded for 25 subjects who were minimally-verbal [14]. These were calculated by summing scores on items 1-9 for each of the two BOSCC segments and averaging the totals. For both scores, higher numbers indicate greater severity of social

impairment.

Pearson correlation coefficients were computed between ADOS CSS SA and the automated eye contact measures (frequency and duration rate) from the ESCS segment and the BOSCC segment (separately). In addition, correlations were computed between the BOSCC SA and the automated eye contact measures (frequency and duration rate) from the BOSCC segment.

BOSCC SA scores for the remaining 41 subjects in the ASD group were not assigned as the subjects had more verbal language than the minimally-verbal level for which the BOSCC is validated, as indicated by the subject having completed an ADOS-2 module 2 or 3, or speaking in flexible phrases or sentences during the BOSCC.

All subjects considered for correlation analysis are a subset of ASD toddle samples. Since they were part of the training set, we use the reduced model that is used for reproducibility analysis for [70] that is trained with 56 subjects, in order to minimize bias. We avoid dropping additional subjects further from this as it would cause shrinking the training set too much and degrade its performance. As a result, subjects included in correlation analysis are partly represented in the sample that the model was trained with. Namely, there are N = 58 (39 trained) subjects for ADOS CSS SA correlation during the BOSCC, N = 45 (25 trained) subjects for ADOS CSS SA correlation during the ESCS, and N = 25 (16 trained) for BOSCC SA correlation analysis.

2.3.6 Baseline: Pose-implicit Convolutional Neural Networks Detector

In this section, we describe the implementation and experimental details of the work of [38] which serves as a baseline model for the proposed eye contact detector.

Model: This model has been designed based on our observation on two major problems with the prior PEEC approach from [43]. First, due to the fact that the method requires head pose estimates (to assign each sample to one of the pose clusters) and eye localizations (to extract features), gaze estimation cannot be performed when landmark detection

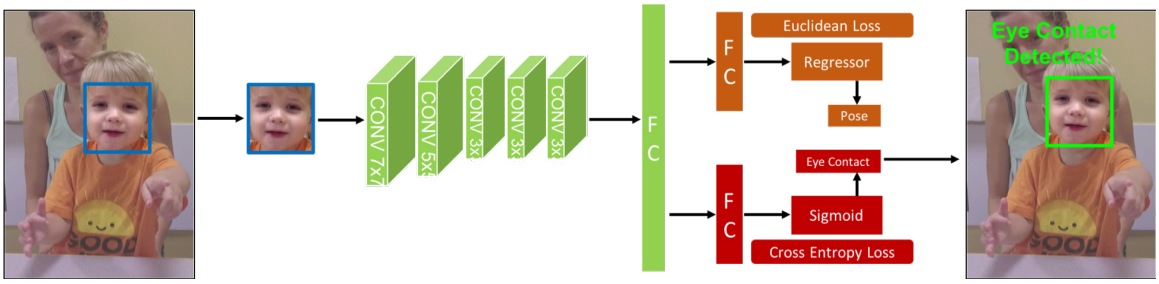


Figure 2.5: **Pose-implicit Convolutional Neural Networks (PiCNN) architecture for eye contact detection.** It utilizes a two stream network structure with 8 layers modeled after AlexNet. One output branch predicts head pose and the other predicts eye contact. The inputs to the network are face bounding boxes. A representative face bounding box is illustrated with a red outline in the image at the left.

fails. Thus any failure along the preprocessing pipeline will result in a missed detection. Facial landmark detection is a more difficult problem than face detection. It is challenging to detect the landmarks reliably when the face is occluded or foreshortened due to out-of-plane rotation, which happens quite frequently during dynamic social interactions with children. Our experiments reveal that, in our dataset, landmarks are successfully found in only 75.56% of the total eye contact frames, whereas faces are detected 97.94% of the time. The second problem with PEEC is its reliance on hand-designed HOG features and a random forest classifier, which are known to be inferior to modern deep neural network classifiers which support end-to-end feature learning.

We therefore proposed a different classification approach based on Convolutional Neural Networks (CNN) [72] that addresses these two issues. This approach learns an image representation that *jointly* predicts head pose and eye contact, instead of being dependent on precomputed head pose and facial landmarks. Since the representation is learned end-to-end (i.e., the input is raw pixels and the output is a binary prediction) there is the opportunity to learn features that are specific to the task. This approach is particularly promising in light of the large-scale social interaction dataset that we have assembled. Our proposed model is called *"Pose-implicit CNN Detector"* or *PiCNN*.

Fig 2.5 illustrates the PiCNN architecture. The input to the network is a rectangle of pixels corresponding to a face bounding box produced by a face detector. The detected

face patch is resized to 227×227. The network has five convolutional and pooling layers, similar to AlexNet [72], but with a smaller filter size (7 by 7) with strides of 2 in the first layer to capture the finer details in the face. Layers 6 to 8 are fully connected. The primary difference between our approach and AlexNet is the presence of two branches in the 7th and 8th fully-connected layers. The upper branch outputs a prediction of the three axes of head rotation (i.e., regression in yaw, pitch, roll) and the lower branch is tasked with the binary classification of eye contact. The weights and parameters are the same and are linked during training in the first 6 layers (5 convolutional and 1 fully connected) and branched in the last 2 layers to facilitate multi-task learning. Thus the prediction of eye contact can benefit from implicitly learning the variability in eye appearance resulting from head pose change. This model achieves the best performance, by a large margin, among all existing methods on our dataset.

In training the PiCNN model, we have ground truth eye contact labels for every frame in the training dataset. This allows us to backpropagate training error through the eye contact detection branch in every training batch. However, we do not have frame-level human annotations of head pose for any frames. We solve this problem by using the IntraFace system as a source of head pose training data. In frames for which head pose estimates from IntraFace are available, we additionally backpropagate training error through the head pose branch. Note that this strategy has the benefit that the network is forced to learn representations for predicting eye contact even when a head pose reference is not available, making it potentially more robust than a sequential approach in which head pose must be computed before detection can be performed. At the same time, our method can take advantage of sparse head pose annotations where they are available and improve the detection performance. During testing time, we apply the PiCNN model in feedforward mode and do not utilize any separate head pose estimates.

Dataset: The data utilized in our experiments comes from four separate studies, conducted at one or more of the following three sites: the Georgia Tech Child Study Lab in

Table 2.2: Summary of the dataset used in the baseline model

age (in months)		gender		diagnosis		ethnicity		protocol		annotation	
	# subjects		# subjects		# subjects		# subjects		# sessions	# frames	minutes
less than 20	8	male	74	TD	50	Caucasian	60	ESCS	30	2,364,773	1,314
$20 \sim 29$	33	female	26	ASD	50	mixed	16	R-ABC	34		
$30 \sim 39$	15					African American	10	v-BOSCC	43		
$40 \sim 49$	7					Asian	6	nv-BOSCC	26		
$50 \sim 59$	4					Hispanic	4	Marcus	23		
$60 \sim 69$	14					unknown	4				
$70 \sim 79$	5										
$80 \sim 89$	5										
greater than 90	9										

Atlanta, GA (GT); the Center for Autism and the Developing Brain in White Plains, NY (CADB); and the Marcus Autism Center in Atlanta, GA (MAC). Descriptive information for the subset of subjects whose data was included in the current analysis and details of the data collection protocol for each study is detailed below, separately for each dataset, and in aggregate in Table 2.2. Sample images from our dataset are illustrated in Fig 2.3.

Generally, all four data collection protocols involved a semi-structured play interaction between an adult and a child, who sat across a small table from each other. The specific protocols chosen were selected based on their prior use in research on social attention and communication in typically developing children and children with autism, and because they have been shown to reliably elicit eye contact from these groups. In all four studies, the examiner interacting with the child wore a pair of commercially-available glasses - Pivothead Kudu - which have an outward-facing camera embedded in the bridge over the nose. By virtue of its placement, this camera reliably captures a close-up image of the child's face as the examiner interacts with the child. The lenses were removed from the glasses to provide the child an unobstructed view of the examiner's eyes. Our prior research indicates that the presence of the glasses does not affect the gaze behavior of children [36].

Research assistants annotated the pivothead videos from each dataset using one of two video-annotation software tools: ELAN (http://tla.mpi.nl/tools/tla- tools/elan/) and INTERACT Mangold (https://www.mangold-international.com). Ground truth coding involved flagging the frame-level onset and offset of each instance of the child making eye contact with the examiner, as captured by looks into the camera. Kappas for frame-level agreement

between pairwise comparisons of the 6 coders ranged from .89 to .94.

Experiment Overview: In this section, we provide experimental evaluations for each of the components of our method. To evaluate the frequency of successful landmark detection (and by extension, head pose estimation) in our dataset, we calculated the percentage of frames labeled as eye contact for which landmarks were detected. Note that this is not a statement about the accuracy of the landmarks, only about their availability. This is meaningful because state-of-the-art methods such as IntraFace will only output landmarks if they are of sufficiently high quality. The percentage of frames with missing landmarks is the portion of the error rate for eye contact detection which is attributable to landmark detection failure. Landmark detection is successful in only 75.56% of frames, in comparison to 97.94% for face detection. This suggests that around a quarter of the errors are due to landmark detection failure, and a negligible amount are due to face detection failure. These findings explain the low recall rate of the two eye contact methods (PEEC and GazeLocking) that rely on landmark detection.

Experimental Setup: We divide the total dataset summarized in Table 2.2 into 5 disjoint train/test splits for 5-fold cross validation, where each split has 80% of training and 20% of testing sessions and subjects included in the training set are not present in the testing set. This ensures that testing is always done on unseen subjects. Each training set split is sampled uniformly across combinations of diagnosis (TD/ASD) and play protocols. For example, we sample 80% for a training set from the TD ESCS group, from the ASD ESCS group, from the TD nv-BOSCC group, from the ASD nv-BOSCC group, and so on, such that different conditions are fairly represented across the five splits. We use the same five folds to train, test and compare the four eye contact detection models – our PiCNN model, PEEC [43], Modified AlexNet [72], and Gaze Locking [45]. With this cross validation approach, we are able to obtain testing results on all of our datasets, and we report the overall performance averaged over all sessions as well as within different populations. Table 2.3 summarizes the results.

On average, the training set in each split initially had 145k positive (eye contact) frames and 1,746k negative (no eye contact) frames, which is highly imbalanced. To overcome the data skewness issue, we resampled the training sets with horizontal flip, slight rotation and color jittering with positive set oversampling and negative set subsampling to make it more balanced at positive:negative = 561k:842k = 4:6 ratio.

Note that our analysis does not include direct performance comparisons to other appearance-based gaze tracking methods such as [48, 50]. The primary reason is that these methods are designed for a different task, accurately mapping the subject's point of gaze on a mobile screen. These methods can identify *where* a subject is looking, but they only know *what* the user is looking at if they have access to ROI masks for the gaze stimulus. In contrast, our method automatically identifies *what* the subject is looking at, but only for the specific gaze event of eye contact. Moreover, the domain of screen viewing is quite different from naturalistic social interactions. This can be seen by comparing Figs 2.2 and 2.3. In particular, [48, 50] rely on landmark localization [73] to extract the eye region from the face. Based on our results, these methods will take an immediate 22.4% miss in recall due to the difficulty of detecting landmarks under challenging conditions. These methods are therefore unlikely to be competitive with PiCNN for our task.

Experimental Results: All four eye contact detection methods output a confidence score between 0 and 1 at each frame, with a higher score indicating increased likelihood of eye contact. We use these per-frame scores to evaluate how well the model is detecting eye contact. Since our dataset has more than 92% negative samples, accuracy ($\frac{tp+tn}{p+n}$) is not a good measure of performance. For example, simply predicting everything as negative will give 92% accuracy but that is not a useful detector. Instead, we report the detector's performance with respect to three metrics, namely F_1, Matthews correlation coefficient (MCC), and the Area Under Curve of Precision Recall curve (AUC-PR).

Both F_1 and MCC are widely used in machine learning as a measure of the quality of binary classifiers. Because the prediction output of the models we used in our analysis

is a real-numbered value instead of a hard binary output, we calculate the F_1 and MCC scores at every cutoff points and report the maximum. Additionally, we also compute the Area Under Curve of Precision Recall curve (AUC-PR). Typically, AUC is computed using Receiver Operating Characteristic (ROC) instead of Precision and Recall curve (PR curve), but AUC of ROC is not sensitive to the uneven class sizes, thus we use the PR curve. Like F_1 score, AUC-PR is maximum at 1 and minimum at 0, but it is an aggregated measure across all prediction cutoffs. Our final results by these criteria are summarized in Table 2.3. Clearly, our PiCNN method outperforms all other methods when evaluated on all datasets as well as on individual groups with different conditions. When the results are broken down into different diagnostic groups, play protocols, gender and ethnicity, in all cases and in all metrics, our PiCNN method achieves the best performance, followed by AlexNet, PEEC and GazeLocking in this order. The standard deviation of the scores is greatest under different play protocol settings (F_1=0.0465, MCC=0.0492, AUC-PR=0.065), and second greatest under different ethnic groups (F_1=0.034, MCC=0.035, AUC-PR=0.0336).

Table 2.3: **Result scores of the baseline model** in five metrics, grouped by different conditions.

		F_1	MCC	AUC-PR	Precision	Recall
All	PiCNN (Ours)	**0.78**	**0.77**	**0.79**	**0.75**	**0.80**
	AlexNet [72]	0.73	0.72	0.75	0.71	0.77
	PEEC [43]	0.63	0.62	0.57	0.67	0.59
	GazeLocking [45]	0.52	0.50	0.48	0.52	0.52
ASD	PiCNN (Ours)	**0.76**	**0.75**	**0.78**	**0.75**	**0.78**
	AlexNet [72]	0.72	0.71	0.74	0.69	0.74
	PEEC [43]	0.64	0.64	0.60	0.68	0.61
	GazeLocking [45]	0.51	0.49	0.49	0.50	0.52
TD	PiCNN (Ours)	**0.79**	**0.78**	**0.78**	**0.74**	**0.83**
	AlexNet [72]	0.75	0.74	0.74	0.69	0.81
	PEEC [43]	0.63	0.62	0.55	0.65	0.62
	GazeLocking [45]	0.54	0.53	0.49	0.54	0.54
ESCS	PiCNN (Ours)	**0.75**	**0.75**	**0.76**	**0.69**	**0.82**
	AlexNet [72]	0.69	0.69	0.72	0.62	0.78
	PEEC [43]	0.57	0.56	0.47	0.55	0.60
	GazeLocking [45]	0.48	0.46	0.41	0.46	0.50
v-BOSCC	PiCNN (Ours)	**0.76**	**0.74**	**0.79**	**0.75**	**0.77**
	AlexNet [72]	0.73	0.71	0.75	0.72	0.74
	PEEC [43]	0.69	0.67	0.67	0.72	0.65
	GazeLocking [45]	0.56	0.53	0.56	0.53	0.60
nv-BOSCC	PiCNN (Ours)	**0.82**	**0.81**	**0.83**	**0.82**	**0.81**
	AlexNet [72]	0.77	0.76	0.78	0.76	0.78
	PEEC [43]	0.71	0.71	0.68	0.76	0.67
	GazeLocking [45]	0.58	0.56	0.55	0.57	0.58
R-ABC	PiCNN (Ours)	**0.77**	**0.77**	**0.71**	**0.72**	**0.84**
	AlexNet [72]	0.72	0.72	0.68	0.66	0.80
	PEEC [43]	0.59	0.59	0.49	0.66	0.54
	GazeLocking [45]	0.52	0.51	0.43	0.53	0.51
Marcus	PiCNN (Ours)	**0.86**	**0.86**	**0.88**	**0.86**	**0.87**
	AlexNet [72]	0.80	0.79	0.83	0.77	0.83
	PEEC [43]	0.54	0.54	0.49	0.67	0.45
	GazeLocking [45]	0.46	0.44	0.41	0.49	0.44
male	PiCNN (Ours)	**0.77**	**0.75**	**0.77**	**0.74**	**0.79**
	AlexNet [72]	0.72	0.71	0.73	0.69	0.76
	PEEC [43]	0.64	0.63	0.57	0.66	0.61
	GazeLocking [45]	0.52	0.50	0.48	0.51	0.53
female	PiCNN (Ours)	**0.78**	**0.78**	**0.80**	**0.76**	**0.81**
	AlexNet [72]	0.74	0.74	0.76	0.70	0.80
	PEEC [43]	0.64	0.63	0.59	0.68	0.60
	GazeLocking [45]	0.53	0.52	0.49	0.54	0.52
Caucasian	PiCNN (Ours)	**0.76**	**0.75**	**0.76**	**0.73**	**0.79**
	AlexNet [72]	0.72	0.71	0.72	0.68	0.76
	PEEC [43]	0.64	0.63	0.57	0.66	0.62
	GazeLocking [45]	0.51	0.50	0.47	0.51	0.52
mixed	PiCNN (Ours)	**0.76**	**0.75**	**0.81**	**0.74**	**0.79**
	AlexNet [72]	0.75	0.74	0.76	0.72	0.79
	PEEC [43]	0.69	0.67	0.66	0.70	0.67
	GazeLocking [45]	0.60	0.58	0.59	0.57	0.63
African American	PiCNN (Ours)	**0.82**	**0.82**	**0.83**	**0.81**	**0.84**
	AlexNet [72]	0.79	0.79	0.79	0.76	0.83
	PEEC [43]	0.62	0.61	0.57	0.68	0.56
	GazeLocking [45]	0.51	0.50	0.48	0.51	0.52
Asian	PiCNN (Ours)	**0.79**	**0.78**	**0.75**	**0.72**	**0.86**
	AlexNet [72]	0.67	0.66	0.63	0.61	0.74
	PEEC [43]	0.65	0.65	0.58	0.66	0.65
	GazeLocking [45]	0.49	0.49	0.41	0.43	0.57
Hispanic	PiCNN (Ours)	**0.73**	**0.73**	**0.78**	**0.68**	**0.79**
	AlexNet [72]	0.70	0.70	0.72	0.66	0.75
	PEEC [43]	0.70	0.69	0.67	0.69	0.71
	GazeLocking [45]	0.58	0.57	0.58	0.55	0.61

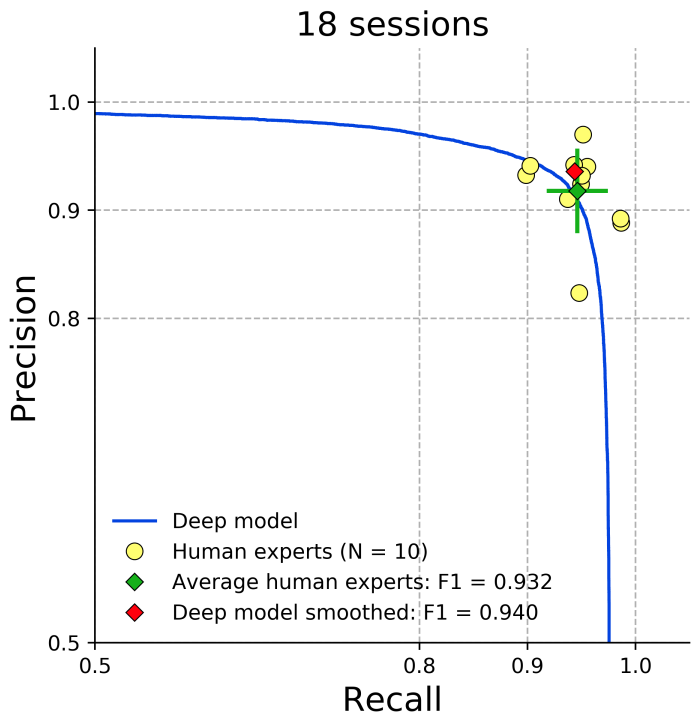


Figure 2.6: **Precision and recall (PR) of deep learning model and human raters on 18 validation sessions.** The blue line is the PR curve for the model, zoomed into the range 0.5-1.0. Improved model PR (red diamond) is obtained by temporally smoothing the model output. The PR for each of the ten expert raters (yellow dots) is obtained by comparing an expert's ratings to the consensus ratings of the other nine experts. Note that the model (red diamond) achieves higher precision than the average of the expert raters (green diamond) for the same recall. The model PR (red diamond) lies within one standard deviation (green error bars) of the mean rater, and both the model and the mean rater have similar F1 scores. Therefore, we conclude that the deep learning model exhibits comparable performance to expert human raters.

2.4 Results

2.4.1 Dataset representation

In t-tests and chi square tests that were run to confirm that subjects included in the validation set are representative of the overall sample, the validation set did not differ from the rest of the sample in terms of diagnostic group (χ = .09, p = .77), gender (χ = 3.62, p = .06), age (t = .49, p = .62), race (χ = 2.70, p = .61), ethnicity (χ = .29, p = .86), or severity of social impairment among the ASD group (t = 1.18, p = .24).

Table 2.4: **Frame-level performance comparisons.** Performance without smoothing (row 3) is reported at the maximum F1 score along the PR curve, with associated PR. In row 5, we replaced AlexNet in [38] with ResNet for a fair comparison.

	F1	Precision	Recall
Mean rater			
ESCS	0.927	0.918	0.935
BOSCC	0.945	0.917	0.975
Combined	0.932	0.918	0.946
Deep model (smoothed)			
ESCS	0.920	0.917	0.924
BOSCC	0.950	0.946	0.954
Combined	0.940	0.936	0.943
Deep model (not smoothed)			
ESCS	0.909	0.898	0.919
BOSCC	0.943	0.940	0.945
Combined	0.930	0.924	0.937
Deep model without transfer learning (smoothed)			
ESCS	0.897	0.893	0.901
BOSCC	0.927	0.935	0.920
Combined	0.916	0.917	0.915
Multi-task learning [38] with ResNet (smoothed)			
ESCS	0.880	0.884	0.876
BOSCC	0.920	0.946	0.896
Combined	0.906	0.924	0.890

2.4.2 Frame-level accuracy

The precision and recall (PR) performance of the deep learning model is illustrated as a blue PR curve in Fig. 2.6. Each yellow dot in the figure gives the PR for one of the ten expert raters. This PR is obtained by comparing an expert's ratings to the consensus ratings of the other nine experts (effectively treating the nine experts' consensus as ground truth). The mean rater (green diamond) is the average of the PRs for the expert raters. The red diamond gives the PR of the model following smoothing (post-processing). Note that it achieves higher precision than the mean rater for the same recall.

Table 2.4 breaks out the performance by dataset and quantifies the benefits of smoothing and transfer learning. Comparing the F1 scores (first column) of the first two large rows

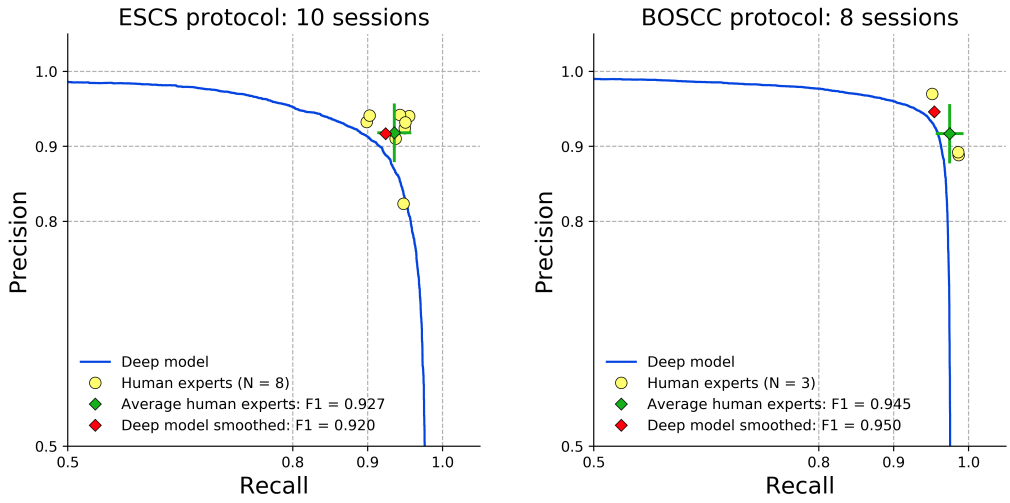


Figure 2.7: **Precision and recall of algorithm output and human ratings shown separately for the ESCS and the BOSCC protocol.** Left figure is based on 10 validation sessions of the ESCS and right figure is based on 8 validation sessions of the BOSCC. ESCS segments were annotated by 8 human experts and BOSCC segments were annotated by 3. One human coder annotated both validation sets. Notations are consistent with those in Fig. 2.6

(divided by horizontal bars) demonstrates the equivalence of the mean rater and deep model performance in all settings. Comparing the 2nd and 3rd large rows, we see that smoothing gives an F1 score increase of 0.01 on the combined dataset. Similarly, removing transfer learning in the 4th row causes a decrease of 0.024 in combined F1. We note that the multi-task learning approach from [38], presented in row 5, causes a drop of 0.034 in combined F1 relative to the transfer learning result in row 2. These results validate the superiority of transfer learning over multi-task learning for this problem. The average precision can be interpreted as the area under the PR curve, and gives an overall measure of the effectiveness of the classifier without the need to select a particular operating point. In Table. 2.4, we report results for specific operating points. Here we provide the average precision for the classifier without performing smoothing: ESCS 0.948, BOSCC 0.959, and Combined 0.956.

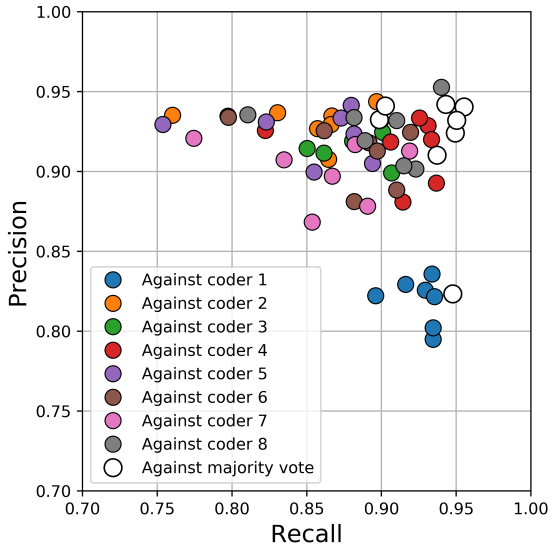


Figure 2.8: **Precision and recall of each human coder (N = 8) under different ground truth on the ESCS validation set.** Color indicates type of ground truth (white: majority vote, other: single human) and each dot denotes one annotator's coding quality measured by the respective ground truth. With the same coding data, the choice of ground truth changes precision and recall.

2.4.3 Reliability with human raters

Inter-rater reliability is measured by Cohen's κ for all pairs of human coders and provides evidence for the reliability of the raters. Using the combined dataset, the average human-human κ is $m_{hh} = 0.888$ (with 0.8 as the standard cut-off for reliability). Comparison of the eye contact detection model with human raters also indicates reliability, with an average human-detector κ of $m_{hd} = 0.891$. This demonstrates that adding the model-based detector as an additional "rater" to the pool of human coders preserves reliability, reinforcing the claim that the model is equivalent to a human expert.

Table. 2.5 gives the summary of the Cohen's κ statistics for the human-human and human-detector comparison, which are also illustrated in Fig. 2.9 and Fig. 2.10 in a box plot.

The hypothesis that the model is equivalent to a human expert can be tested statistically using two one-sided tests. For the combined dataset, we are able to reject $H_0 : m_{hd} -$

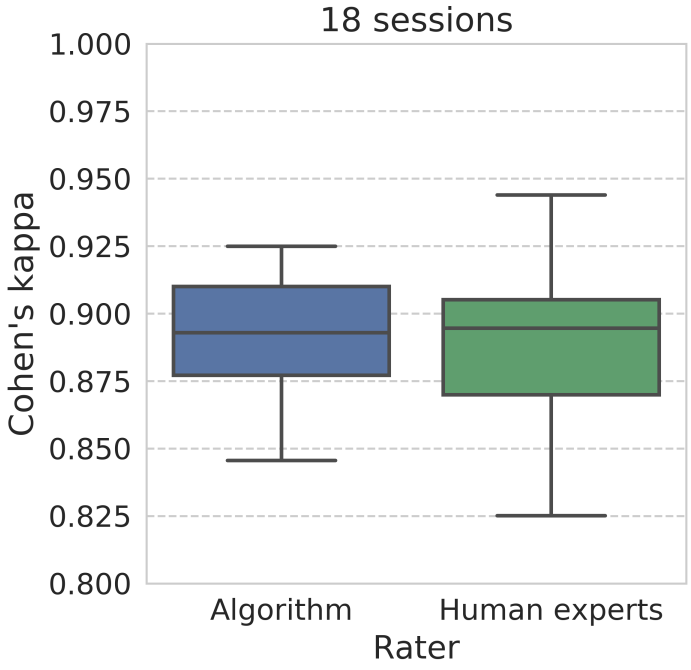


Figure 2.9: **Pairwise Cohen's kappa distributions among all human pairs and human-algorithm pairs**, represented as box plot. Generally, kappa scores above 0.8 is considered an almost perfect agreement. On 18 validation sessions annotated by 10 human experts, agreements among humans and agreements between each human and algorithm are similar in terms of kappa values.

$m_{hh}¡-\Delta$ at p = .05 and accept H_1 that the algorithm is as reliable as human annotators, with the equivalence boundary Δ as low as .025, which is the standard deviation of the human κ's. In the standard two one-sided tests, sample means from the two groups are compared for both sides of inequalities. However, we are only interested in testing if the detector is not less reliable as human. Therefore, the null and alternative hypotheses tested in this analysis is the following. As shown in Table. 2.6, we are able to reject the null hypothesis H_0 at p = .05 for all cases.

$$H_0: \ m_{hd} - m_{hh} < -\Delta$$

$$H_1: \ -\Delta < m_{hd} - m_{hh}$$

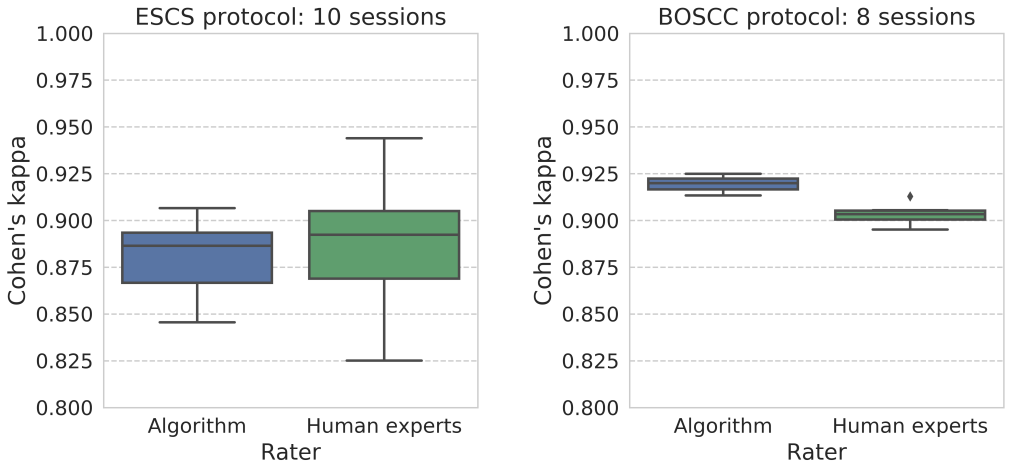


Figure 2.10: **Pairwise Cohen's kappa distributions among all human pairs and human-algorithm pairs**, represented as box plot separately for the ESCS (left) and the BOSCC (right) protocol.

Table 2.5: Cohen's κ reliability statistics

	κ, human-human pairs	κ, human-detector pairs
ESCS		
Min	.825	.846
Max	.944	.907
Average	.886	.880
BOSCC		
Min	.895	.913
Max	.913	.925
Average	.903	.919
Combined		
Min	.825	.846
Max	.944	.925
Average	.888	.891

Table 2.6: **Reliability equivalence test statistics.** All test cases are statistically significant at p = .05. Δ = 0.025

	m_{hh}	m_{hd}	p for $H_0 : m_{hd} - m_{hh} < -\Delta$
ESCS	.886	.880	$< .03$
BOSCC	.903	.919	$< .005$
Combined	.888	.891	$< .005$

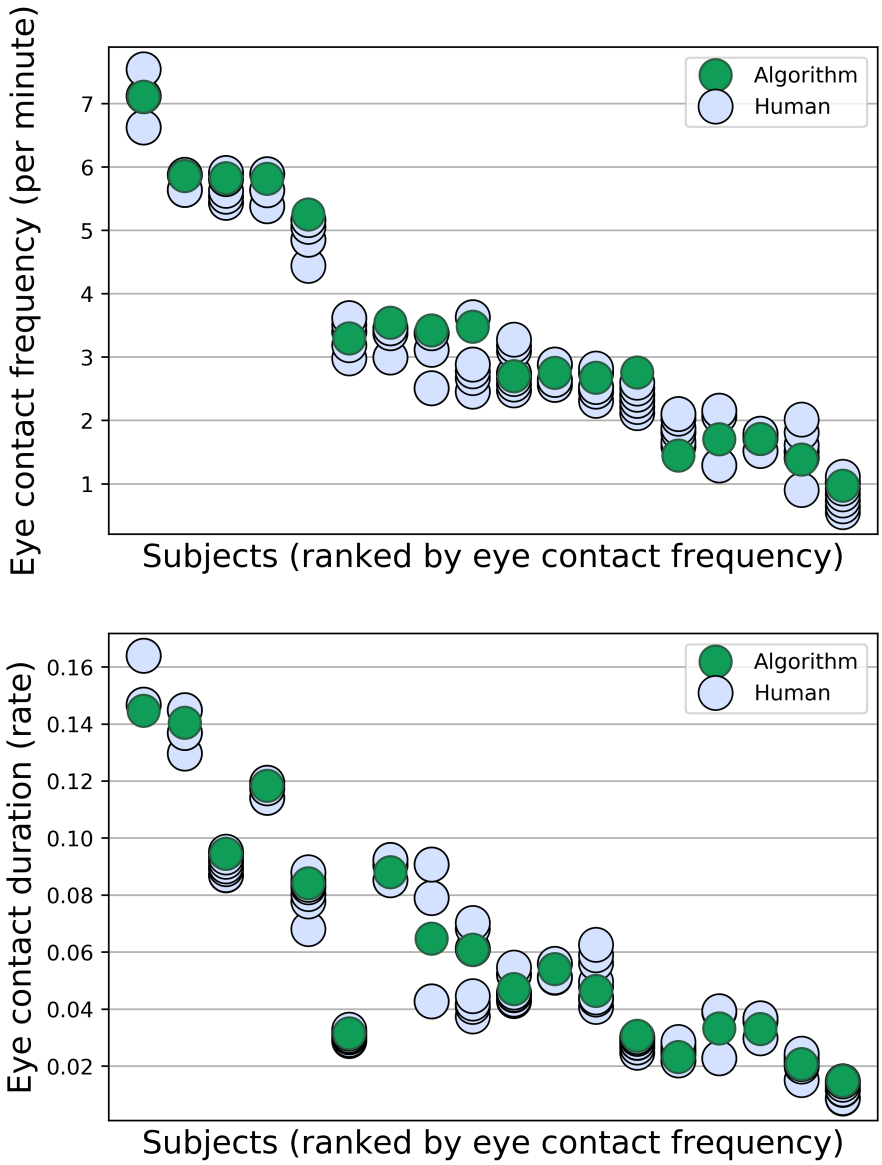


Figure 2.11: **Eye contact frequency (top) and duration (bottom) coded by human and algorithm for 18 validation subjects.** Frequency is how many times a subject made eye contact per minute, and duration is the length of eye contact normalized over the administration time. In both figures algorithm's estimate (green) is within the distribution of manual annotation (light blue). X axis is sorted by the subject's eye contact frequency such that two figures are aligned in the x axis. Subjects may make short eye contacts frequently (e.g., 3th, 6th subject) which can be also measured by the algorithm.

Table 2.7: **Original and reproduced statistical tests of [70].** EC eye contact. Independent 2-group Mann-Whitney U test is used in cross-group analysis and Wilcoxon Signed-Rank test is used for within-group analysis. (*) Statistically significant at p= .05.

	Based on manual coding (p)	Based on automated coding (p)
Cross-group analysis: TD (N = 38) vs. ASD (N = 21)		
TD more EC during toy inactive than ASD	.001*	.007*
TD more EC in child possession than ASD	.01*	.02*
TD more EC during toy active than ASD	.06	.06
Within-group analysis: within TD (N = 38)		
More EC during toy inactive than active	$< .001$*	$< .001$*
More EC in examiner possession than in child possession	$> .1$	$> .1$
Within-group analysis: within ASD (N = 21)		
More EC during toy inactive than active	$< .001$*	.007*
More EC in examiner possession than in child possession	$> .1$	$> .1$

2.4.4 Reproducibility of prior studies

Fig. 2.11 provides a visualization of the eye contact frequency and duration rate for each subject, comparing the detection model (green circle) and human raters (gray circles). The figure demonstrates good qualitative agreement across subjects, with the model estimates consistently falling within the range defined by the human coders. Note that some subjects are harder to rate, resulting in a greater spread of measures. We replicated the hypotheses tests for significance from the prior studies using the automated coding results. Automated findings were *identical* to those obtained from human coding. Table 2.7 summarizes the findings for study [70], with significance for the effect of context on eye contact duration and frequency, but not for the the effects of time or time-context interaction. Table 2.8 and Fig. 2.12 give the results for study [37]. Note that all of the subjects used in this analysis were excluded from the model training set.

Table 2.8: **Original and reproduced statistical tests of [37].** Percent duration and rate of eye contact during interactive play and conversation across the two time points were compared in 2 (context: play, conversation) by 2 (time: T1, T2) ANOVAs (N = 15). (*) Statistically significant at p= .05.

	Based on manual coding (p)	Based on automated coding (p)
Duration of eye contact		
Effect of context	$< .001$*	$< .001$*
Effect of time	$> .9$	.84
Interaction between time and context	$> .7$	$> .7$
Frequency of eye contact		
Effect of context	$< .001$*	$< .001$*
Effect of time	$> .9$	.82
Interaction between time and context	$> .7$	$> .7$

Figure 2.12: **Average duration of eye contact during conversation and interactive play in child and adolescent sample (N = 15)**, measured at time 1 and time 2, based on human coding (left) and automated coding (right). Error bars denote standard error.

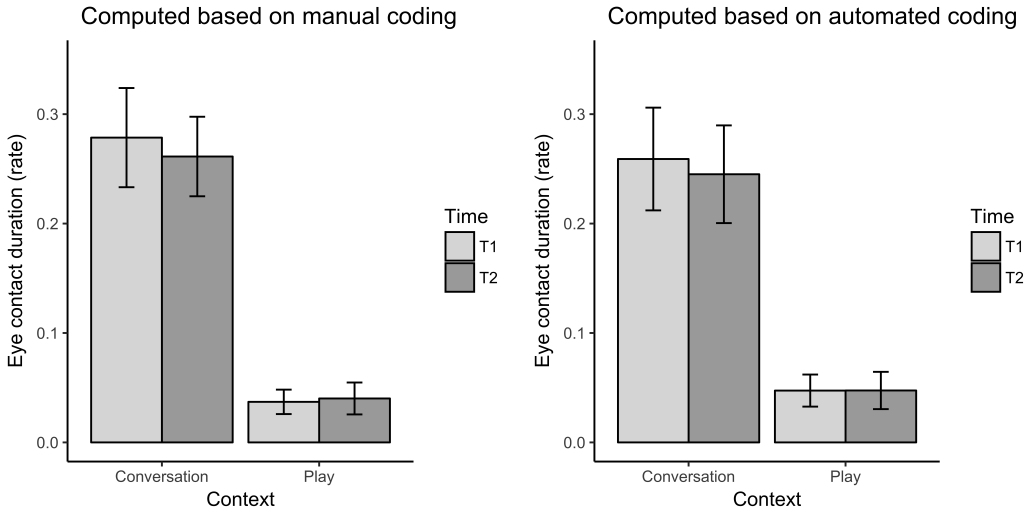
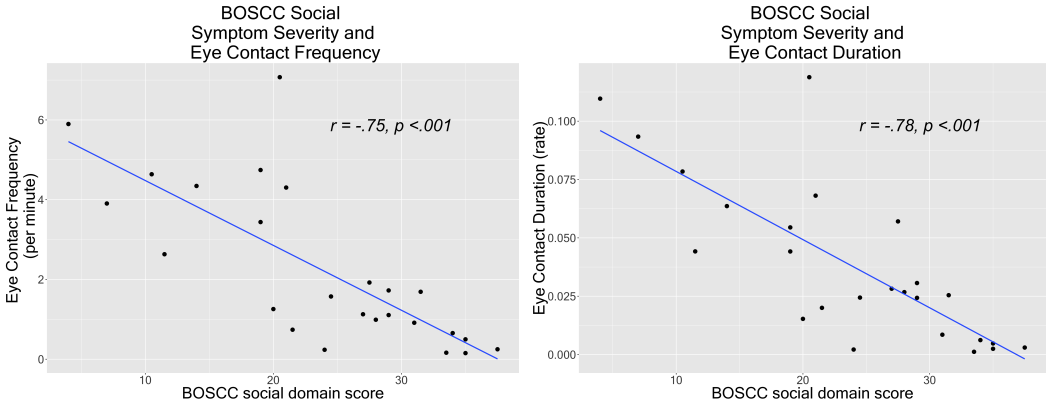


Figure 2.13: **Correlation between automatically measured eye contact and the severity of social impairment** during the BOSCC among subjects with ASD (N = 25). In both frequency (left) and duration (right) are strongly negatively correlated with the severity score.



2.4.5 Correlation analysis

For individuals with ASD, ADOS CSS SA demonstrated a weak relation with frequency and duration of automatically measured direct gaze during the ESCS (N = 45. frequency: r = -.41, $p < .01$; duration: r = -.36, $p < .05$) and during the BOSCC (N = 58. frequency: r = -.26, $p < .05$; duration: r = -.29, $p < .05$), which was mostly driven by a small number of subjects with low social affect severity scores. Severity of overall social symptoms during the BOSCC (BOSCC SA) demonstrated a strong correlation with both frequency and duration, illustrated in Fig. 2.13 (N = 25. frequency: r = -.75, $p < .001$; duration: r = -.78. $p < .001$).

2.5 Conclusion

This chapter provides multiple, converging sources of evidence that a deep learning model can detect moments of eye contact in point-of-view camera video with *reliability equivalent* to expert human raters. Our findings validate the use of automated analysis as a substitute for human coding in application domains ranging from autism [70, 37] to job interviews [74, 75], and enable the *scalable* measurement of eye contact during face-to-face interactions. We evaluated our method on a diverse dataset containing children and adolescents from both typical and ASD populations. We believe these are the first findings of equivalence in precision and recall between an automated eye contact detector and human raters.

We now summarize the evidence supporting equivalence between the deep model and expert raters. First, frame-level PR performance for the combined dataset, illustrated in Fig. 2.6, demonstrates that the smoothed deep model result (red diamond) achieves higher precision than the mean human rater (green diamond) at the same recall. Note also that both the red diamond and the unsmoothed PR curve (in blue) lie within one standard deviation (green error bars) of the mean rater. Second, reliability analysis provides additional evidence to complement the PR analysis results. Treating the deep model as an additional rater results in human-detector reliability of 0.891, while human-human reliability was 0.888. The distribution of kappas is illustrated in Fig. 2.9. The hypothesis that the deep model is as reliable as human raters was examined using two one-sided statistical tests and found to hold with an equivalence threshold as low as 0.025. The third source of evidence comes from correlation analysis between eye contact frequency and duration and ASD symptom severity. We would expect increased severity to be negatively correlated with rates of eye contact, and this is born out in Fig. 2.13 which illustrates a strong negative correlation between automatically derived eye contact measures and the BOSCC Social Affect (BOSCC SA) scores. Weaker correlation was found between the derived measures

and the ADOS CSS SA scores. The fourth source of evidence comes from a reproducibility study which asks whether the findings from two recently-published studies of eye contact in autism will continue to hold if automatically-derived measures are used in place of expert ratings. Table. 2.7, Table. 2.8 and Fig. 2.12 demonstrate that all findings hold, providing direct evidence for the feasibility of using automatically-derived eye contact measures in developmental studies.

An additional contribution of this work is to explore the relative merits of transfer learning and multi-task learning approaches in learning effective models for eye contact. As shown in Table 2.4, the proposed transfer learning approach that first learns 3D head pose and gaze and then learns eye contact is superior to both the previously proposed multi-task approach [38] that simultaneously learns head pose and eye contact and a baseline method that simply learns eye contact without learning 3D head pose models. Note that if we were determining eye contact using an external room camera, then the estimation of head pose and gaze angle in 3D would be vitally important to determine where the subject is looking in space. By locating the camera on the subject, we reduce this global 3D estimation task to the much simpler task of assessing gaze relative to the coordinate frame of the camera. Nonetheless, the need to make angular determinations may explain why pretraining on an explicit 3D estimation task leads to improved performance.

We briefly review relevant prior work to place our contribution in context. First, a variety of recent works have demonstrated the feasibility of using deep learning to achieve expert level analysis of biomedical data [76] for detecting and classifying clinical conditions such as diabetic retinopathy [77], skin cancer [78], malignant mammographic lesions [79], bone fracture [80], and atrial fibrillation [81]. In contrast, only a few prior works have explored the automated analysis of social behaviors in clinical contexts such as autism. Early works on automatically analyzing social behaviors [82, 11, 83, 84] predated the development of deep learning technology and did not address the issue of expert level performance. Prior work on automatically measuring "response to name" behaviors [85, 86]

included both ASD and typical samples and assessed the agreement with expert human raters, but did not address naturalistic face-to-face social interactions.

Marinoiu et al. [87] use deep learning models to analyze interactions between children with ASD and a robot therapist, but don't address expert-level performance. Note that none of these prior works addressed the assessment of eye contact. A final line of related work uses machine learning tools to improve the usability of conventional eye tracking technology by improving robustness to head movements and minimizing the need for calibration [48, 88, 65, 89]. While works such as [48] use deep learning to analyze gaze, they focus on the case of gaze to screens or displays. We believe this is the first work to demonstrate human level performance in automatically assessing a social behavior in a naturalistic face-to-face interaction context.

While the focus of this work has been on the assessment of eye contact in interactions between an unencumbered child and an examiner, our technology could also be applied to the analysis of face-to-face interactions between adults in which each subject is wearing video recording glasses. Additional applications in clinical and social psychology [29, 30, 31, 32, 33] could potentially benefit from this approach. Moreover, these methods can also support the development social intelligence for robots, enabling them to interact naturally with people using nonverbal social signals. Mutual gaze and joint attention are found to have a critical role in conversation, narration, collaboration, and manipulation tasks between humans and robots [90, 91]. We will release the trained models and software from this work to facilitate such future work.

CHAPTER 3

HEAD POSE-BASED ATTENTION SHIFT MEASRUEMENT

3.1 Introduction

The problem of capturing and analyzing social behaviors during naturalistic interactions is an important and challenging task with a broad range of applications in automated behavior analysis and social robotics. For example, the use of sensors and machine learning methods to analyze social behaviors has emerged recently as a promising technology for understanding and treating developmental conditions such as autism [92, 93]. Moreover, in the area of human-robot interaction, there is a long-standing interest in creating social robots with nonverbal communication capabilities [94, 95]. While motion capture technology can be used to record social behavior, it requires the use of professional actors as marker-based methods are too invasive to capture spontaneous naturalistic interactions between multiple people. This is particularly true in the case of *children's* social behaviors. There has been a limited amount of prior work on the analysis of children's behavior from video [43, 96, 97]. These works have tended to focus on facial expression analysis or the detection of specific behaviors such as eye contact. While several software packages exist for tracking facial landmarks [98, 99], facial expressions are only one element of social behavior. In particular, facial expressions are coordinated with shifts of attention, and attention in turn requires the coordination of *head movement* with the eyes. Head movement provides additional nonverbal communication cues, such head nods and shakes for "yes" and "no." In addition, the 3D location and pose of the head identifies the portion of the scene that the person is facing and is likely to be attending to. It follows that the ability to track head pose and localize heads in 3D is a key capability for social behavior capture and analysis.

While there have been a variety of prior works on head tracking from video [100, 101,

102], few of these methods are designed to work with multiple video cameras, and as a consequence they are limited to relative head pose and cannot localize the head in a 3D room coordinate system. There have been a few works on multicamera head tracking [103, 104] along with works that focus on more general multicamera reconstruction which can recover 3D head location [105, 106]. Unfortunately, these methods are not suitable for the large-scale capture of children's social interactions due to the expense and complexity of their multi-camera setup. It is commonplace to record assessment and therapy sessions with children using a single room camera, but is not practical to capture with the large number of cameras needed for dense reconstruction.

As an alternative, we have developed a practical and effective approach to capturing children's social behaviors which combines a single room camera with a wearable camera worn on an adult social partner. This setup reflects the fact that measurement of a child's behavior frequently occurs via interactions with an adult, such as a clinician, therapist, teacher, or caregiver. We call this setup *face plus context* because the head-worn camera of the adult examiner provides almost continuous capture of the child's face and head, while the room camera provides access to the social context, and supports localization in a 3D room coordinate system.

We present a novel multi-camera system for 3D head tracking and localization which is suited to the face plus context scenario. We combine continuous tracking and calibration of the head-worn camera with 3D localization and pose estimation of all heads in the social scene. Our system uses state-of-the-art methods for face tracking and head pose estimation combined with multi-target tracking to provide 3D localization and disambiguate identity in the case where there are multiple people present.[1] Our system automatically tracks all heads in the scene and reconstructs the pattern of social interaction between the participants based on head movement. This is a first step towards a more comprehensive 3D social capture system which will incorporate gestures and gaze shifts in addition. This work

[1]For example, it is common for very young children to sit on a parent's lab during an assessment, with the result that the parent's head becomes a distractor for the task of tracking the child.

makes the following contributions:

- We present the first 3D multi-head tracking and capture system for the *face plus context* measurement scenario, which is designed to be applicable to a wide range of child assessment scenarios including behavioral screening, therapy, evaluation, and skill training contexts.

- We provide the first experimental results for 3D head tracking of children and their adult social partners during naturalistic face-to-face social interactions.

- We demonstrate that head shifts detected using our 3D head tracking approach are a useful step in detecting gaze shifts, based on experiments including both typically developing children and children with autism. Our results are promising in light of the difficulty of deploying standard gaze tracking technology in naturalistic social scenarios.

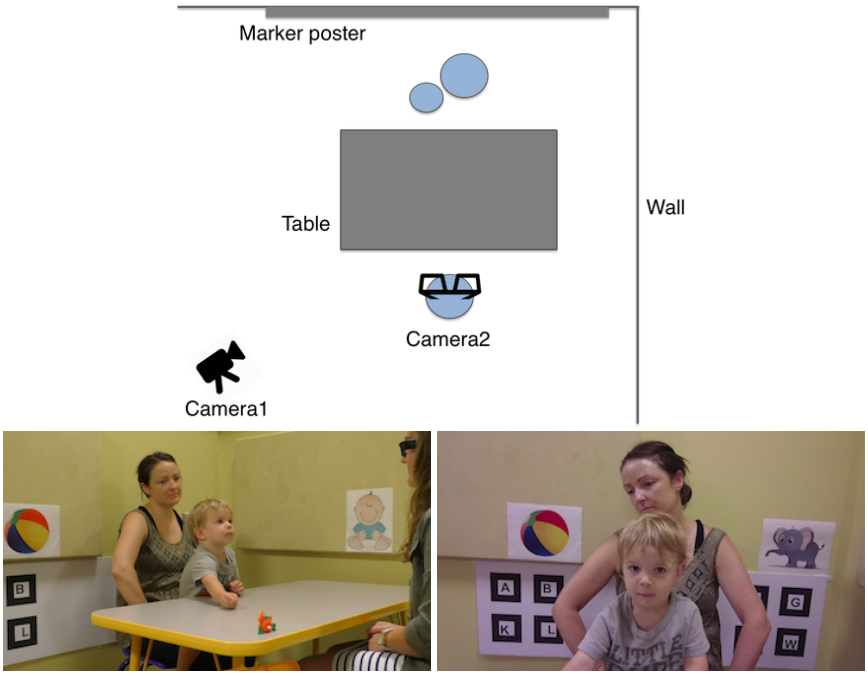


Figure 3.1: Our setup "face plus context" and sample images from each camera.

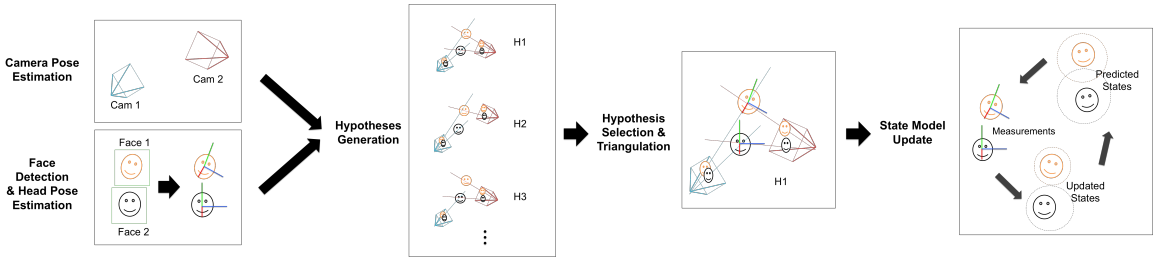


Figure 3.2: **System overview.** First, each camera pose is estimated based on the patterns placed on the wall (Sec. 3.3.2). Also, face is detected and head pose w.r.t. each camera is estimated using facial landmark alignments (Sec. 3.3.3). Finally, the most likely combination is used to update head state models (Sec. 3.3.4).

3.2 Related Work

There are three main areas of prior work which are relevant to this chapter: 3D head pose estimation, head pose-based social sensing, and video-based measurement of children's behaviors.

3D Head Tracking: There is a large body of work on face localization and tracking from video [102]. Single camera tracking systems can estimate the head pose relative to the camera and track the scale of the face, but are unable to localize heads in 3D reliably. This includes methods based on facial landmarks [98, 107] and other alignment methods [108, 109], as well as full 3D head models [110, 111]. If two or more cameras are used, then it is possible to recover the complete 3D location and pose of the head. Since children frequently lean towards the objects and people that they are interacting with, the 3D head location is a valuable cue for social understanding. There have been several prior works on multi-camera head tracking [112, 103, 104, 113, 114]. One area of prior work, of which [104] is representative, use multiple cameras to estimate where a user is looking in a smart environment application. The goal is to determine which of a discrete set of gaze targets is being attended to. Other works in meeting room understanding, described below in more detail, use head pose (along with audio cues in some cases) to understand patterns of conversation, floor holding, and other communicative acts. None of these works produce continuous measurements of 3D head location and pose, a capability provided

by our approach which creates the possibility for more fine-grained measures of social behavior. Our use of multiple hypothesis tracking (MHT) is related to the work of Benfold and Reid [115], which demonstrated the ability to track multiple pedestrians using a single camera in an outdoor surveillance application and determine where they were looking. In contrast, our MHT approach can produce full 3D estimates of head locations and pose over time from multiple cameras.

Head Pose-Based Social Behavior Sensing: Prior works on social sensing have utilized estimates of head pose as a cue for social attention, although none of these works have addressed children's behavior. One representative problem is understanding patterns of conversation and attention between adults during business meetings [113, 114]. In comparison, our task requires the consideration of a broader range of gaze targets, as children will interact with toys as well as their social partners. Moreover, these works assume that participants don't approach each other closely and can therefore be tracked without a multiple hypothesis tracking framework. In our case, when children sit on their parent's lap or approach the examiner it creates a more challenging tracking problem. Another line of work [116, 117] uses head orientation to understand social group formation and detect specific types of social interactions, such as conversational group detection based on f-formation theory from social psychology. In contrast, we are interested in detecting specific behaviors of interest, such as a shift in gaze based on head movement, not in classifying the type of interaction.

Our use of a wearable camera connects us to other works that study social attention from a first person vision perspective. The closest work is [118], which shows that patterns of social interaction within a group can be classified based on the change in the head poses of the group over time, as captured by a wearable camera. This work assumes that there are many people looking at the same gaze targets, and requires that the targets be visible to the wearable camera. In contrast, our case is primarily a dyadic interaction, and often the child's gaze targets are not visible in the examiner's camera (requiring the use of a room

camera). Another representative work identifies social "hotspots" that arise when many people, all wearing head-mounted cameras, look at the same location [119]. This approach does not handle behaviors like eye contact, and it would require instrumenting the child, which would limit the applicability of the method considerably. One point that we have in common with all of these prior works is an assumption, often called "center bias," that head orientation is a powerful indicator of one's direction of attention [120]. Our experimental results confirm that this bias is a useful cue in analyzing children's attention as well.

Vision-based Measures of Children's Social Behavior: Other works have demonstrated the ability to measure aspects of a child's social behavior using vision. In dyadic face-to-face interactions, it has been shown that facial expression analysis can be used to discover synchrony between an infant and their caregiver [97]. In earlier work, we demonstrated the ability to detect moments of eye contact using a wearable camera [43]. We go beyond these works by addressing children's attention to objects in addition to faces. A small number of works address activity recognition in children from video [121, 122], and while this has no direct bearing our work it is part of the broader story for behavior capture using sensors.

3.3 Methods

Our sensing approach for the *face plus context* capture scenario utilizes two cameras to maximize coverage of the child's behavior while facilitating automated measurement. The setup is illustrated in Fig. 3.1. The first camera is statically mounted and records the scene context. The second camera is concealed in a pair of glasses worn by the examiner and captures the child's face (see Sec. 3.3.1 for details). Our analysis pipeline is illustrated in Fig. 4.4. The first step is to calibrate the two cameras so they can be combined to localize the heads. Since the wearable camera is in constant motion, it is continuously-calibrated using AR Tags (see Sec. 3.3.2). Separately, faces are detected and relative head pose is estimated from each camera (see Sec. 3.3.3). The set of detections is processed by

a multiple hypothesis tracking algorithm (see Sec. 3.3.4) which maintains the identity of each subject over time and fuses both camera views to produce 6 DOF location and pose for each head. Head movements are then analyzed to predict attention to elements of the scene (see Sec. 3.3.5). We now describe these elements in more detail.

3.3.1 Face Plus Context Setup

Our goal is to support the capture of a child's behaviors in the context of interaction across a tabletop, which is a standard assessment paradigm in psychology. The setup is illustrated in Fig. 3.1. Young children may sit on a parent's lap, with the result that the parent is frequently visible in both cameras. The tabletop serves as a convenient surface for toy manipulation and also helps to constrain the child's movement. In this setting, the examiner administers a set of play protocols by interacting with the child using a set of toys. In the case of assessing children with autism, the protocols are designed to elicit social behaviors such as eye contact or pointing, which are key elements of joint attention. This procedure makes it possible to tap a variety of social behaviors in a well-defined context.

We utilize two cameras to capture the child's social behaviors. The first camera is mounted on a tripod and is placed at an angle that covers the table and the people in the scene. It provides capture of the entire social scene from a fixed vantage point, ensuring that the child and any toys they are interacting with will be visible at all times. It produces a lower resolution image of the child, but is still useful for localizing the child in 3D. The second camera is inconspicuously located in a pair of glasses (Pivothead SMART) worn by the examiner. The glasses can be filled with a prescription or the lenses can be removed to provide an unobstructed view. Since the examiner naturally maintains an orientation to the child at all times, this camera provides continuous high resolution capture of the child's face. Moreover, the position of this camera facilitates the detection of eye contact via the method of [43]. In order to support continuous calibration of the head-worn camera, a single poster board with black and white patterns (AR Tags) printed on it is attached to one

wall so that it is visible to both cameras. Note that the addition of the wearable camera and poster board are all that is needed to convert a standard psychology assessment into our face plus context scenario.

3.3.2 Camera Pose Estimation

In order to reliably estimate camera pose from different views, we use a marker-based pose estimation approach using an open-source Augmented Reality (AR) software called ARToolKit (https://artoolkit.org). This utilizes a square pattern (marker) designed so that it can be detected easily and its 6-DOF pose relative to the camera can be computed reliably. Since the pattern can be occluded by the participants, we use 8 square marker patterns, printed on a single foam board which is mounted on the wall. Fig. 3.3 illustrates this step. Pose can be estimated from one or more detected markers. This method requires knowledge of the camera's intrinsic parameters. We perform checkerboard calibration of the intrinsics at the start of the session. For the wearable camera, calibration can be done once and reused across different sessions. The room camera intrinsics need to be recalibrated if the camera zoom or other settings change.

We considered using a structure-from-motion approach and we evaluated several different SLAM methods [123, 124] on our dataset. Since the videos captured by the POV camera contain abrupt motions, a narrow camera field-of-view, and a dynamic scene, these structure-from-motion methods had difficulties in identifying stable features, and primarily tracked points on the moving humans, which is not useful for camera pose estimation.

According to the reported accuracy of marker-based pose estimation by ARToolKit [125], our setup is within the range of permissible error (2 degrees). In future work, we could explore the use of bundle adjustment with the initial pose estimate to improve the accuracy further.

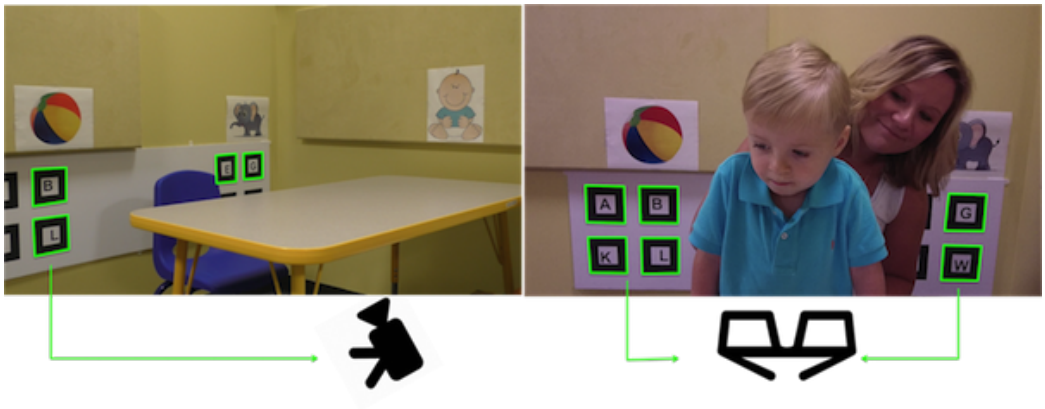


Figure 3.3: **Camera pose estimation.** Green boxes show detected markers used for pose estimation of a room camera and a wearable camera.

3.3.3 Face Detection and Head Pose Estimation

In each view, faces are detected using the Omron OKAO library (https://www.omron.com). For each detected face, we use IntraFace [98] to find and track facial landmarks and we use the Perspective-n-Point algorithm to estimate head pose relative to the camera coordinate frame. The result is a set of 5-DOF measurements, 3 for the head pose R_w^{face} in (3.1), and 2 for the face ray $X(\lambda)$ in (3.2), since the depth is unknown:

$$R_w^{face} = R_w^c \times R_c^{face} \tag{3.1}$$

gives the head pose R_w^{face} in world coordinates based on the estimated camera pose R_w^c and the estimated face pose w.r.t. camera R_c^{face}.

$$\begin{aligned} P^+ &= P^T(PP^T)^{-1} \\ X(\lambda) &= P^+x + \lambda C \end{aligned} \tag{3.2}$$

gives the ray $X(\lambda)$ passing through the face center (i.e. face ray), based on the camera center C, the camera projection matrix P, and the image pixel location of the face center x.

3.3.4 Data Association and Tracking

One of the challenges in analyzing social interactions is to track each person's head consistently without confusing one participant for another (i.e. ID-switch errors). We now describe the multi-target tracking framework we developed to solve this problem. We start by describing the 6-DOF head state model which defines the state space for tracking and supports the fusion of observations from both cameras over time.

Head State Model

For each head, we define a state

$$\begin{aligned} S = (&x, y, z, \dot{x}, \dot{y}, \dot{z}, \ddot{x}, \ddot{y}, \ddot{z}, \\ &q_1, q_2, q_3, q_4, r_1, r_2, r_3, \dot{r_1}, \dot{r_2}, \dot{r_3})^T, \end{aligned} \tag{3.3}$$

where x, y, z is location, $\dot{x}, \dot{y}, \dot{z}$ is velocity, $\ddot{x}, \ddot{y}, \ddot{z}$ is acceleration, q_1, q_2, q_3, q_4 is rotation in quaternion, r_1, r_2, r_3 is angular velocity, and $\dot{r_1}, \dot{r_2}, \dot{r_3}$ is angular acceleration. Then, each state is tracked with an Extended Kalman Filter following the process update model in (3.4) and measurement update model in (3.5). The process update model is

$$\begin{aligned} S_t^- &= f(S_{t-1}, w_{t-1}) \\ P_t^- &= A_t P_{t-1} A_t^T + W_t Q_{t-1} W_t^T \\ A_t &= \frac{\partial f}{\partial s} \\ W_t &= \frac{\partial f}{\partial w}, \end{aligned} \tag{3.4}$$

where S_t^- is predicted state, P_t^- is predicted state covariance, w is process noise, Q is process noise covariance, f is the function that projects the positional data and angular velocity linearly, except for q that is updated through a quaternion multiplication with d,

the difference caused by angular velocity. The measurement update model is

$$\begin{aligned} h(S_t, v_t) &= (x, y, z, \frac{q}{|q|}) + v_t \\ K_t &= P_t^- H_t^T (H_t P_t^- H_t^T + V_t R_t V_t^T)^{-1} \\ S_t &= S_t^- + K_t(m_t - h(S_t^-, 0)) \\ P_t &= (I - K_t H) P_t^- \\ H_t &= \frac{\partial h}{\partial s} \\ V_t &= \frac{\partial h}{\partial v}, \end{aligned} \tag{3.5}$$

where h is the function that extracts position and rotation from the state vector, v is measurement noise, K is Kalman gain, R is measurement noise covariance, m is taken measurement, P is updated covariance matrix. Note that we let x, y, z, q be measurables, and how this measurements are generated and associated is described in detail in the following section.

Data Association

Initially, per view and per detected face, we have a measurement of 5 degrees of freedom,

$$m = (X(\lambda), q1, q2, q3, q4), \tag{3.6}$$

where q is quaternion of R_w^{face} in (1) and $X(\lambda)$ is from (2). For instance, if two faces are detected in camera 1 and two faces are detected in camera 2, there should be four individual m's at that moment. Then, we define a cost function $C(m, S)$ between a pair of m's and a state S, considering both geometric and appearance constraints as follows:

$$C(m, S) = w_g \times C_g(m, S) + w_a \times C_a(face, S), \tag{3.7}$$

where w_g and w_a are weight parameters, C_g is Mahalanobis distance between m and S, which is calculated using P^- in (4), and C_a is classification score of a detected face to the state. In the very first frame, each state is initialized by assigning a face to S_{parent} as -1 or to S_{child} as 1, and a linear regressor is trained online subsequently. With this cost function, all m's are assigned to a state S independently per camera. Meanwhile, an m with a cost above a certain threshold is discarded to remove spurious detections, incorrect pose estimations, etc. Note that this matching is done per view, such that an m from a given view is associated with one state S exclusively (or discarded), but each state can have multiple m's from different views. As a result, if one state is assigned two m's, they are triangulated to obtain 3D position, and the rotation is interpolated using the slerp algorithm. If only one m is assigned to a state, the closest 3D point on the line is selected. This is the m_t used in the measurement update model (3.5). To reflect the confidence of the final measurement, the measurement noise covariance R in (3.5) is reduced as the number of m's used increases.

3.3.5 3D Scene Estimation

In addition to 3D head tracking, we also estimate the location of the table and the toy. We estimate table pose by using the same approach as in camera pose estimation. In the beginning of each session, a known square pattern is put on the table for a few seconds. This is sufficient to retrieve table pose as the table does not move throughout session. Additionally, the play protocol we utilized incorporates a set of toys presented at one location on the table (Fig. 3.5), for which the estimated table pose can be used as well. Additional work could be done to refine the toy locations further, for example using pose estimation from a 3D model [126] or triangulation with a custom toy-object detector.

3.3.6 Data

In this section, we describe the data used in the chapter and how it was collected and processed.

Participants: Participants were recruited and data was collected at Georgia Tech (GT) and Weill Cornell Medicine (WC). Our dataset consists of eight sessions from typically developing (TD) children and eight sessions from children with autism. Eight TD children (3 female) with no known diagnosis of social, developmental, or communication delays were recruited at GT via community advertising and a parent mailing. Eight children with a diagnosis of ASD (4 females) were recruited by WC. The diagnosis of ASD was confirmed prior to participation by a licensed clinician. TD participants were between 20 and 36 months of age (mean age = 30.8 months), and ASD participants were between 32 and 60 months of age (mean age = 43.8 months). All participants completed play-based assessments during a single visit.

Play Protocol: All participants completed a modified version of the Early Social Communication Scales [12], a semi-structured, examiner-directed assessment of nonverbal communication skills in young children. The child is seated, sometimes on a caregiver's lap, at a small table across from the examiner. The examiner presents several different toys and activities to the child, selected because of their potential to elicit joint attention (using gaze and gestures to share the experience of objects or events with a social partner) and requesting (using nonverbal behaviors to elicit aid in obtaining objects or events). The toys include: a) three small wind-up mechanical toys, b) three hand-operated toys, c) a small car and a ball that will roll easily across the table, d) a book with large distinct pictures on its pages, and d) colorful posters positioned on the walls to the left, right and behind the child. The ESCS administration takes about 15-25 minutes to complete.

The present analysis focused on the object spectacle toys, which include three unique wind-up toys and three hand-operated toys, including a trapeze monkey, a balloon, and a spintop. For each of the six object spectacle tasks, the examiner places one of the six toys in the corner of the table to her left, activates the toy for about 5-10 seconds, then allows the toy to remain inactive for about 5-10 seconds. Per the scoring rules in the ESCS manual, any time the child shifts their gaze from the active toy to the examiner's eyes and then

back to the toy, they are credited with engaging in initiating joint attention. When the toy becomes inactive, similar shifts of attention are credited as initiating behavior regulation (i.e., requesting).

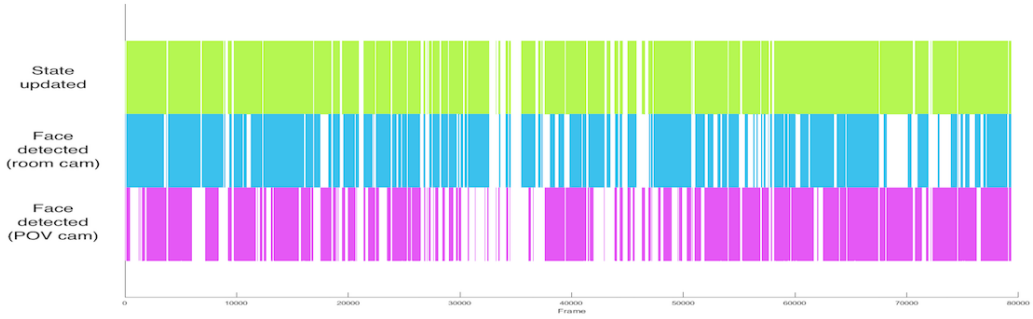


Figure 3.4: State update frequency and face detection frequency.

Annotation: Videos of the ESCS assessments from the room camera and the POV camera were used for manual coding by trained raters to identify each moment when the child was looking at the toy or making eye contact with the examiner. The start and the end of each spectacle toy presentation was coded as well. Coders used Mangold International's Interact annotation software (https://www.mangold-international.com/en/software/interact) to identify these moments and mark the onsets and offsets at the video frame level.

3.4 Results

In this section, we report quantitative and qualitative evaluations of our approach.

3.4.1 Head Tracking Statistics

We first evaluated the performance of our system by calculating overall tracking statistics. We ran the tracker on 15-25 minute-long sessions of 16, which resulted in 1,152,600 total frames. Among these, the child head state model has been updated with new measurements 87.6% of the time, meaning that the child's face was successfully detected in 87.6% of the frames. Within this detection rate, 80.1% of face detection is from the room camera

Figure 3.5: **Social signals captured by our system.** Three axes of red, green, and blue represent child's head pose. First column shows the 3D head trajectory during 4 seconds in the direction of the arrow. Row 1: Wind-up toy is presented and the child is requesting the examiner to give it to him by making eye contact. Row 2: Examiner is pointing to a poster and the child is following. Row 3: Examiner is choosing a toy and the child is peeking over the table.

and 67.5% is from the POV camera. The reasons why it is lower in the POV camera are: A) limited field of view, B) camera viewpoint change, and C) motion blur. While B) naturally occurs following wearer's head movements, A) and C) can be improved with the advances of wearable camera technology. Despite these challenges, the inclusion of a wearable camera adds great value to the system overall. This can be seen in Fig. 3.4, which identifies frames in which face detection failed for the room camera (vertical white lines in the middle row) but succeeded for the POV camera (last row) leading to greatly increased state updates (top row is more dense than either face detection row alone). This finding confirms the effectiveness of our face plus context setup.

3.4.2 Qualitative Results

Children show a wide range of head movements in the course of ESCS. This can be observed by visualizing the 3D head trajectories and their reprojections on the input videos. As shown in Fig. 3.5, children use a broad range of head and body movements to communicate with people and achieve their goals. As a consequence, our videos comprise a rich,

dynamic, and densely-sampled (60 Hz frame rate) database of social motion in a variety of contexts.

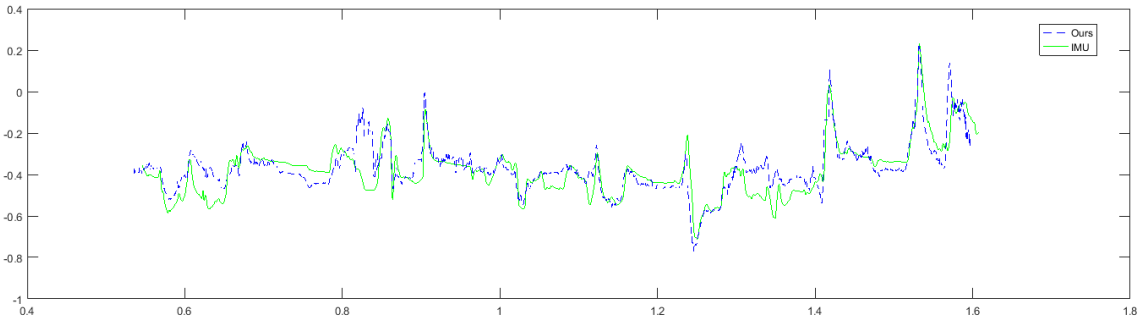


Figure 3.6: IMU vs. video-based head tracker.

To further evaluate our method, we collected videos during a short ESCS session in which an IMU sensor was tightly-attached to head. The results from this experiment (see Fig. 3.6) indicate that the head motions measured by the IMU sensor and follow a very similar pattern to our head tracking system (mean error = 0.12, variance = 0.08, in radians).

3.4.3 Detection of Gaze Shift for Joint Attention

To assess the viability of using head pose estimates to measure gaze shifts, we devised a simple detector for moments when the child looks at a spectacle toy and subsequently makes an eye contact with the examiner. This type of gaze shift is known as initiating joint attention (IJA). IJA is more frequent in TD than in ASD and is correlated with language outcomes. During ESCS object spectacle tasks, toys are presented at a known 3D location at the corner of the table. We trained a classifier to analyze the head tracking data and detect the child's shifts of attention from the toy to the examiner. Specifically, we identified time segments in which the child's yaw angle decreased (segments from cyan up to red point of Fig. 3.7-2), and created features using the distances from child's head z-vector to the examiner and toy (green and red curve of Fig. 3.7-1). Each segment was labeled as an IJA shift or not based on the ground truth annotations. 10% of the 1665 total segments are IJA shifts. We trained a binary SVM classifier using 20% of the samples, obtaining the ROC

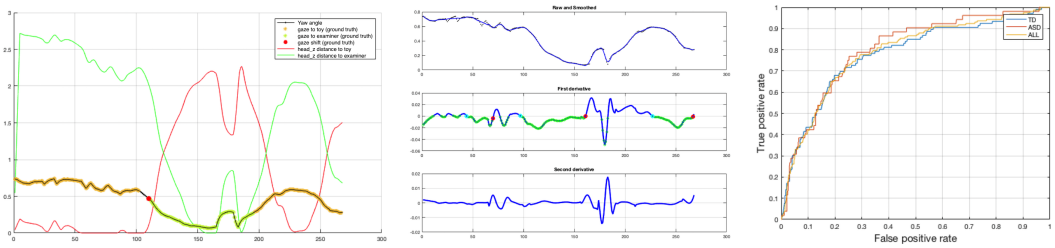


Figure 3.7: **Gaze shift detection.** Left figure shows how measurements change over time when there is a gaze shift from toy to examiner. Middle figure shows how a segment is selected at testing time for gaze shift detection. Right figure is an ROC curve showing our gaze detector's performance.

curve in Fig. 3.7-3. The detector performed equally well on both diagnostic groups (AUC score 0.78 for TD, 0.8 for ASD). Moreover, the predicted total number of gaze shifts per subject is correlated with the ground truth gaze shift counts (p-values: TD < 0.005, ASD < 0.0077, All < 0.0007).

3.4.4 3D Attention Map

We developed a method for generating a 3D attention density map to support additional visualization and understanding of our collected measures. We begin by creating a 3D volumetric scalar field to represent the gaze density at any 3D point in the interaction space. We utilize a gaze model similar to [127], wherein the gaze vector is assumed to lie in a cone-shaped distribution emanating from the center of two eyes, capturing the uncertainty in head pose and eye gaze. The head pose estimate from our system directly gives a 3D vector (Z axis of the head coordinate system) emanating from the center of the eyes in the direction of the front of the head. The uncertainty in the actual gaze direction with respect to the head pose estimate is represented by a Gaussian distribution on a plane normal to the head direction vector.

For computational feasibility, the volumetric scalar field is discretized into 3D voxels (which are analogous to pixels in a 2D image). For speed and memory efficiency, we use a 512^3 voxel array to represent the entire the 3D interaction space. Each voxel stores a scalar score representing the likelihood of gaze at that voxel. The scores are initialized to zero

and recursively updated for all head pose estimates. With our cone-shaped gaze distribution model, each voxel's gaze likelihood score is updated according to the voxel's 3D location with respect to the gaze distribution. The scores are simply aggregated across head pose estimates to produce the final cumulative 3D gaze likelihood. Given this representation, we can compute a heat-map on any 3D surface by extracting a slice through the attention density map. An advantage of this 3D volumetric approach is that it can accommodate any other arbitrary gaze model.

Fig. 3.8 gives an example of the application of our attention model to a sequence in which a child is reaching for a toy. The cone-shaped gaze distribution in 3D space and its reprojection on the image are shown. Fig. 3.9 illustrates the process of cumulative map generation over a period of time and the final cumulative volumetric map sliced along the table surface, reprojected on a room camera image, and color-coded in heat map color scheme. This example corresponds to a toy presentation period, explaining the high density at the corner where the toy is observed, and a smaller peak in the vicinity of the examiner.

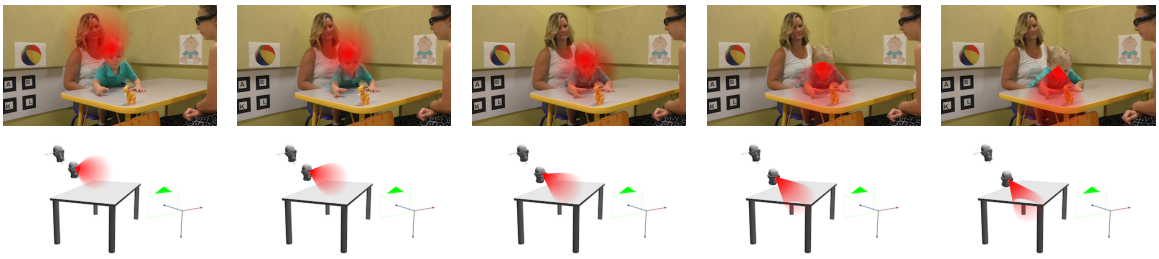


Figure 3.8: **Visualization of our gaze distribution model when a child reaching for a toy.** Bottom row shows it in the reconstructed 3D space and top row shows it by reprojecting the model on actual image.

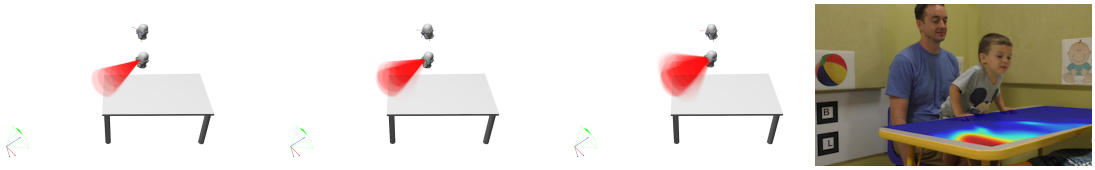


Figure 3.9: **Cumulative attention map during a toy presentation period.** The cumulative map generation process and the final heat map along the table plane.

3.5 Conclusion

We have presented a novel method for automatically capturing children's head motion in face-to-face naturalistic social interactions. Our flexible camera setup and automated tracking framework makes our system especially suitable for the large-scale capture of children's social interactions. Our method has been successfully applied to 16 sessions that include typically developing children and children with autism, during naturalistic play interactions with an adult examiner. Our experimental results demonstrate that our 3D head tracking approach is effective in measuring children's social behavior, and we present promising results for detecting gaze shifts based on head motion.

CHAPTER 4

GENERALIZED ATTENTION TARGET DETECTION

Figure 4.1: **Visual attention target detection over time.** We propose to solve the problem of identifying gaze targets in video. The goal of this problem is to predict the location of visually attended region (green dot) in every frame, given a track of an individual's head (green box). It includes the cases where such target is out of frame (row-col: 1-2, 1-3, 2-1), in which case the model should correctly infer its absence.

4.1 Introduction

Gaze behavior is a critically-important aspect of human social behavior, visual navigation, and interaction with the 3D environment [20, 128]. While monitor-based and wearable eye trackers are widely-available, they are not sufficient to support the large-scale collection of naturalistic gaze data in contexts such as face-to-face social interactions or object manipulation in 3D environments. Wearable eye trackers are burdensome to participants and bring issues of calibration, compliance, cost, and battery life.

Recent works have demonstrated the ability to measure gaze behavior directly from video, with the potential to greatly increase the scalability of naturalistic gaze measurement. A key step in this direction was the work by Recasens et al. [129], which demonstrated the ability to detect the attention target of each person within a single image. This approach was extended in [130] to handle out-of-frame gaze targets, and other related works include [131, 132, 133]. This approach is attractive because it can leverage head pose features, as well

as the saliency of potential gaze targets, in order to resolve ambiguities in gaze estimation. A key limitation is the restriction to working on single images.

This paper develops a *spatiotemporal* approach to gaze target prediction which models the temporal dynamics of gaze from video data. Fig 4.1 illustrates our goal: For each person in each video frame we estimate where they are looking, including the correct treatment of out-of-frame gaze targets. The ability to model gaze from videos is crucial, because it is the shifts in gaze over time that define key aspects of social behavior [21] and characterize human task performance [128, 134]. Furthermore, this approach has the benefit of linking gaze estimation to the broader tasks of action recognition and dynamic visual scene understanding.

An alternative to the dynamic prediction of gaze targets is to directly classify specific categories of gaze behaviors from video [135, 136, 137, 138, 139]. This approach treats gaze analysis as an action detection problem, for actions such as mutual gaze [135, 136, 137] or shared attention to an object [138]. While these methods have the advantage of leveraging holistic visual cues in detecting complex gaze behaviors of interest, they are limited by the need to pre-specify and label the target behaviors. In contrast, our approach of predicting gaze targets gives more flexibility in modeling domain-specific gaze behaviors, such as the assessments of social gaze used in autism research [12, 140].

A key challenge in tackling the dynamic estimation of gaze targets in video is the lack of suitable datasets containing ground truth gaze annotations in the context of rich, real-world examples of complex time-varying gaze behaviors. We address this challenge by introducing the VideoAttentionTarget dataset, which contains 1,331 video sequences of annotated dynamic gaze tracks of people in diverse situations.

Our approach to spatiotemporal gaze target prediction has two parts. First, we present a novel spatial reasoning architecture to improve the accuracy of target localization. The architecture is composed of a scene convolutional layer that is regulated by the head convolutional layer via an attention mechanism [141], such that the model focuses on the scene

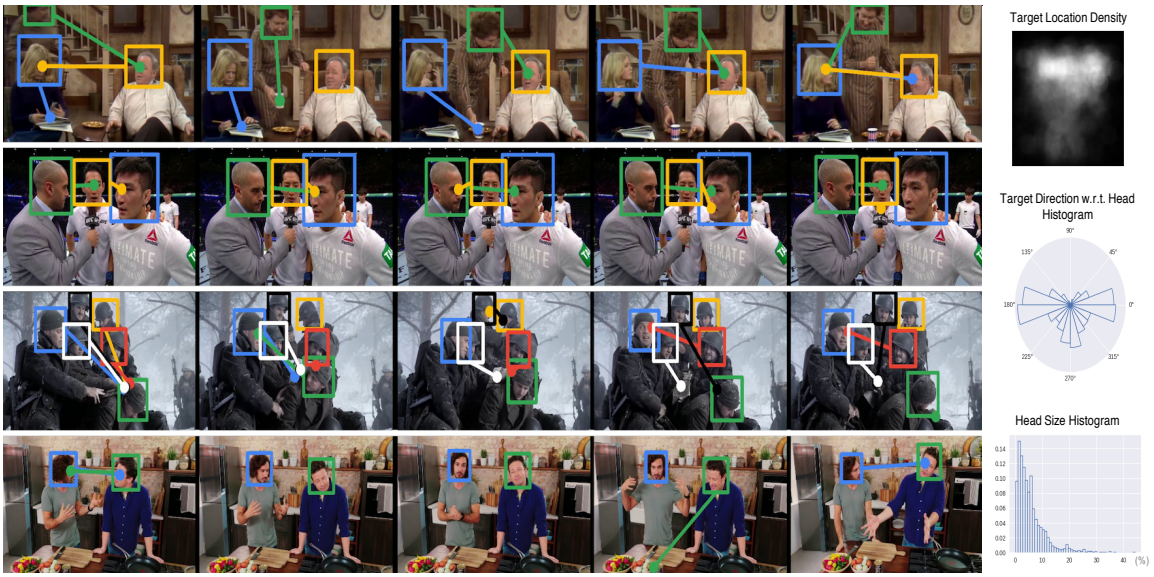
region that the head is oriented to. The spatial module improves the state-of-the-art result on the GazeFollow benchmark (image, within-field-of-view target only) by a considerable margin. Second, we extend the model in the temporal dimension through the addition of ConvLSTM networks. This model outperforms multiple baselines on our novel VideoAttentionTarget dataset. The software, models and dataset will be made freely-available for research purposes.

We further demonstrate the value of our architecture by using the predicted heatmap from our model for social gaze recognition tasks. Specifically, we experimented on two tasks; 1. Automated behavioral coding of social gaze of young children in clinical assessment setting. 2. Detection of shared attention in social scenes. In the first experiment, our heatmap features were found to be the most effective among multiple baselines for attention shift detection. In the second experiment, our approach achieved state-of-the-art performance on the VideoCoAtt dataset [138]. Both results validate the feasibility and effectiveness of leveraging our gaze target prediction model for gaze behavior recognition tasks. In summary, we make the following contributions:

- A novel spatio-temporal deep learning architecture that learns to predict dynamic gaze targets in video
- A new VideoAttentionTarget dataset, containing dense annotations of attention targets with complex patterns of gaze behavior
- Demonstration that our model's predicted attention map can achieve state-of-the art results on VideoCoAtt and two social gaze behavior recognition tasks

4.2 Related Work

We organize the related work into three areas: gaze target prediction, gaze behavior recognition, and applications to social gaze analysis. Our focus is gaze target prediction, but



(a) Example frames and annotations (b) Statistics

Figure 4.2: **Overview of novel *VideoAttentionTarget* dataset** (a) Example sequences illustrating the per-frame annotations of each person (bounding box) and their corresponding gaze target (solid dot). (b) Annotation statistics: top - annotated gaze target location distribution in image coordinates, middle - histogram of directions of gaze targets relative to the head center, bottom - histogram of head sizes measured as the ratio of the bounding box area to the frame size.

we also provide results for behavior recognition in a social gaze setting (see Secs. 4.4.3 and 4.4.4).

Gaze Target Prediction: One key distinction between previous works on gaze target prediction is whether the attention target is located in a 2D image [129, 131, 132, 130, 133] or 3D space [142, 143, 144, 145, 146]. Our work addresses the 2D case. The closest works to ours are [129, 133, 132, 130, 131], and we discuss each in detail. Authors of [129] were among the first to demonstrate how a deep model can learn to find the gaze target in the image. Saran et al. [132] adapt the method of [129] to a human-robot interaction task. They collect a novel dataset with a robot but do not report results on the standard benchmark datasets. Chong et al. [130] extends the approach of [129] to address out-of-frame gaze targets and presents architectural improvements that lead to improved performance on the GazeFollow dataset. In [133], they enhance [129] by considering body pose. A key

difference between [133, 132, 130, 129] and our approach is that we explicitly model the gaze behavior over time and report results for gaze target prediction in video.

Our problem formulation and network architecture are most closely-related to [130]. In addition to our focus on temporal gaze prediction, three other key differences with [130] are 1) that we do not supervise with gaze angles; 2) therefore we greatly simplify the training process; and 3) we present an improved architecture. In terms of architecture, 1) we use head features to regulate the spatial pooling of the scene image via an attention mechanism; 2) we use a head location map instead of one-hot position vector; and 3) we use deconvolutions instead of a grid output to predict a fine-grained heatmap. Our experiments show that these innovations result in improved performance on GazeFollow (i.e., for static images, see Table 4.2) and on our novel video attention dataset (see Table 4.3).

Recasens et al. [131] shares our goal of inferring gaze targets from video. In contrast to our work, they address the case where the gaze target is primarily visible at a later point in time, after the camera pans or there is a shot change. While movies commonly include such indirect gaze targets, they are rare in the social behavior analysis tasks that motivate this work (see Fig. 4.9). Our work is complementary to [131], in that it leverages the temporal dynamics of gaze to infer within-frame gaze targets.

Several works address the inference of 3D gaze targets [142, 143, 146]. In this setting, the identification of an out-of-frame gaze target can be made by relying on certain assumptions about the scene such as the target object's location or its motion. The 3D gaze target may also be inferred in a joint learning framework such as the simultaneous inference of 3D attention, intention and task [144] or people's location, gaze direction and target location [145].

Gaze Behavior Recognition: An alternative to inferring the target gaze location is to directly infer a gaze-related behavior of interest. For example, several approaches have been developed to detect if two people are looking at each other [135, 136, 137], or to detect if more than two people are looking at a common target [147, 138]. Recently, Fan

et al. [139] propose a new problem of recognizing atomic-level gaze behavior when human gaze interactions are categorized into six classes such as avert, refer, and follow.

In contrast to approaches that directly infer gaze behavior, our method provides a dense mid-level representation of attention for each person in a video. Thus our approach is complementary to these works, and we demonstrate in Secs. 4.4.3 and 4.4.4 that our gaze representation has utility for gaze behavior classification.

Social Gaze Detection in Clinical Settings: One motivation for our work is the opportunity for automated measurements of gaze behavior to inform research and clinical practice in understanding and treating developmental conditions such as autism [92, 148]. In this setting, automated analysis can remove the burden of laborious gaze coding that is commonplace in autism research, and enable a more fine-grained analysis of gaze behavior in clinical populations. Prior work in this area has leveraged the ability to analyze head orientation [149, 67, 150] to infer children's attention and have developed solutions for specific settings [151, 152, 153]. Prior works have also addressed the detection of eye contact and mutual gaze in the context of dementia care [154, 155] and autism [43, 38]. Other work has analyzed mutual gaze in group interactions for inferring rapport [156]. In contrast to these works, our focus is to first develop a general approach to gaze target identification in video, and then explore its utility in estimating clinically-important social behaviors during face-to-face interactions between an adult examiner and a child. We believe are the first to present results (in Sec. 4.4.3) for automatically detecting clinically-meaningful social gaze shifts without a wearable camera or an eye tracker.

4.3 Methods

4.3.1 VideoAttentionTarget Dataset

In this section we describe our novel *VideoAttentionTarget* dataset that was created specifically for the task of video gaze target modeling. Some example frames, annotations and statistics of the dataset are shown in Fig 4.2.

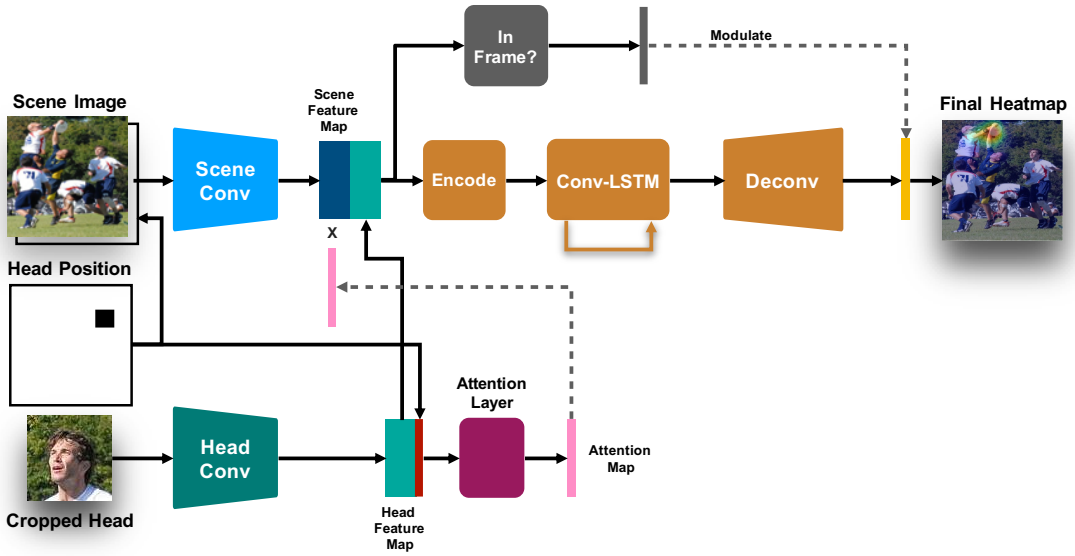


Figure 4.3: **Spatiotemporal architecture for gaze prediction.** It consists of a head conditioning branch which regulates the main scene branch using an attention mechanism. A recurrent module generates a heatmap that is modulated by a scalar, which quantifies whether the gaze target is in-frame. Displayed is an example of in-frame gaze from the GazeFollow dataset.

In order to ensure that our dataset reflects the natural diversity of gaze behavior, we gathered videos from various sources including interviews, sitcoms, reality shows, and movie clips, all of which were available on YouTube. Videos from 50 different shows were selected. From each source video, we extracted short clips that contain dynamic gaze behavior without scene transitions (shot cuts). The length of the clips varies between 2-20 seconds at 25 fps.

For each clip, annotators first labeled tracks of head bounding boxes for each person in each frame. This resulted in 1,331 tracks comprising 164,541 frame-level bounding boxes. In the second pass, the annotators labeled the gaze target as a point in each frame for each annotated person. They also had the option to mark if the target was located outside the video frame (including the case where the subject was looking at the camera). This produced 109,574 in-frame gaze targets and 54,967 out-of-frame gaze annotations. All frames in all clips were annotated using custom software by a team of four annotators, with each frame annotated once.

A testing set was constructed by holding out approximately 20% of the annotations (10 shows, 298 tracks, 31,978 gaze annotations), ensuring no overlap of shows between the train and test splits. In order to characterize the variability in human annotations of gaze targets, two annotators additionally labelled each of the frames in the testing set for samples that were not annotated by that particular annotator, resulting in three independent sets of annotations for all testing frames.

4.3.2 Spatiotemporal Gaze Architecture

Our architecture is composed of three mains parts. A **head conditioning branch**, a **main scene branch** and a **recurrent attention prediction module**. An illustration of the architecture is shown in Fig. 4.3.

Head Conditioning Branch The head conditioning branch computes a head feature map from the crop of the head of the person of interest in the image. The "Head Conv" part of the network is a ResNet-50 [64] followed by an additional residual layer and an average pooling layer. A binary image of the head position, with black pixels designating the head bounding box and white pixels on the rest of the image, is reduced using three successive max pooling operations and flattened. We found that the binary image encoded the location and relative depth of the head in the scene more effectively than the position encoding used in previous works. The head feature map is concatenated with this head position feature. An attention map is then computed by passing these two concatenated features through a fully-connected layer which we call the "Attention Layer".

Main Scene Branch A scene feature map is computed using the "Scene Conv" part of this branch of the network, which is identical to the "Head Conv" network previously described. This scene feature map is multiplied by the attention map computed by the head conditioning branch. This enables the model to learn to pay more attention to the scene features that are more likely to be attended to, based on the properties of the head. In

comparison to [130], our approach results in earlier fusion of the scene and head information. The head feature map is additionally concatenated to the weighted scene feature map. Finally, the concatenated features are encoded using two convolutional layers in the "Encode" module.

Recurrent Attention Prediction Module After encoding, the model integrates temporal information from a sequence of frames using a convolutional Long Short-Term Memory network [157], which designated as "Conv-LSTM" in Fig. 4.3. A deconvolutional network comprised of four deconvolution layers, designated as the "Deconv" module, upsamples the features computed by the convolutional LSTM into a full-sized feature map. We found that this approach yields finer details than the grid-based map used in [130].

Heatmap Modulation The full-sized feature map is then modulated by a scalar which quantifies whether the person's focus of attention is located inside or outside the frame, with higher values indicating out-of-frame attention. This "out-of-frame" scalar is computed from the scene branch before encoding, by the "In Frame?" module in Fig. 4.3, which consists of two convolutional layers followed by a fully-connected layer. The modulation is performed by an element-wise subtraction of the "out-of-frame" scalar from the normalized full-sized feature map, followed by clipping the heatmap values to 0. This yields the final heatmap which quantifies the location and intensity of the predicted attention target in the frame. In Fig. 4.3 we overlay the final heatmap on the input image for visualization.

Implementation Details We implemented our model in PyTorch. The input to the model is resized to 224×224 and normalized. The Attention Layer outputs 7×7 spatial soft-attention weights. The Encode and In Frame modules use 1×1 conv layers. The ConvLSTM module uses four ConvLSTM layers with 3×3 kernels, which produces the heatmap of size 64×64. Further model specifications are available in the supplementary document.

For supervision, we place a Gaussian weight (σ=3) around the center of the target to

create the ground truth heatmap. *Heatmap Loss* is computed using MSE loss if the target is in frame per ground truth. The *In Frame Loss* is computed with binary cross entropy loss. The sum of the two losses is used for backpropagation.

Training is performed in a two-step process. First, the model is globally trained on the GazeFollow dataset until convergence. Second, it is subsequently trained on the VideoAttentionTarget dataset, while freezing the Scene Conv and Head Conv modules to prevent overfitting. We used random flip, color jitter and crop augmentations.

4.3.3 Baseline: Joint Learning of Gaze and Saliency

This section briefly describes the image-based model for attention target detection which is used as a baseline method in our experiment. This work has been published as [130].

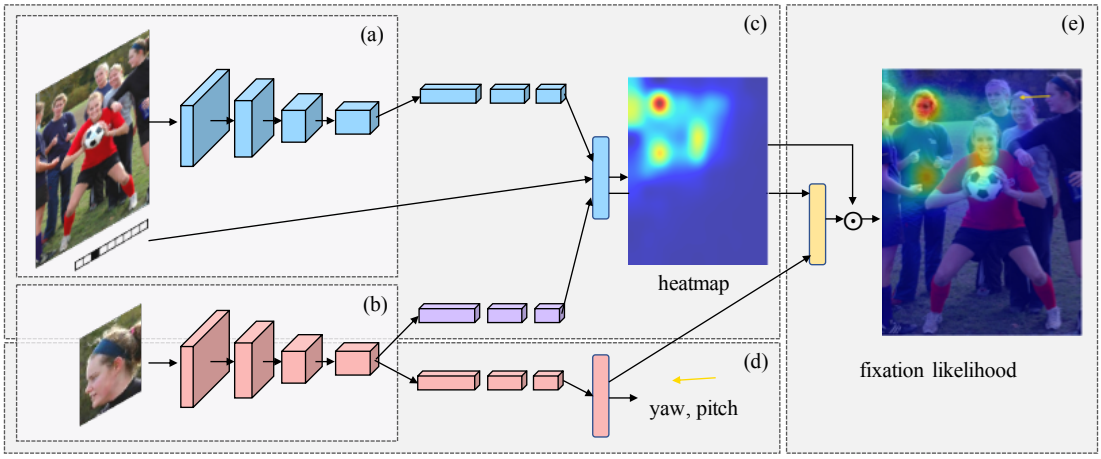


Figure 4.4: **Overview of the architecture.** Full scene image, a person's face location whose visual attention we want to predict, and the corresponding close-up face image is provided as input. Scene and face images go through separate convolutional layers in such a way that (a) (b) and (c) contribute to person-centric saliency, and (b) and (d) contribute to gaze angle prediction. In the very last layer, the final feature vectors for these two tasks are combined to estimate how likely the person is actually fixating at a gaze target within the observable scene.

Model

The inputs given to the model are the entire image, the subject's cropped face and the location of the face of the subject whose attention we want to estimate. The two images are

resized to 227×227 so that the face can be observed in higher resolution by the network. Face position is available in terms of the (x, y) full image coordinates. These coordinates are quantized into a 13×13 grid and then flattened to a 169 dimensional 1-hot vector.

The model consists of two convolutional (conv) pathways: a face pathway (Figure 4.4-d) and a scene pathway (Figure 4.4-c). ResNet 50 [158] is used as a backbone network for the conv pathways (Figure 4.4-a and b). Specifically we use all conv layers of ResNet50 for each of our conv pathways. After each ResNet50 block we add three conv layers (1x1, followed by 3x3, followed by 1x1) with ReLu and batch norm - with stride 1 and no padding. The blue conv layers represented in (c) have filter depth of 512, 128 and 1 respectively. The purple and red conv layers after the face pathway (represented in (c) and (d)) have filter depth of 512, 128 and 16. These conv layers serve to reduce the dimensionality of the features extracted by the ResNet50 backbone networks.

In the face pathway, the feature vector computed with the face input image goes through a fully connected layer to predict the gaze angle represented using yaw and pitch intrinsic Euler angles. In the scene pathway, the feature vectors extracted from the whole image as well as from the face image are concatenated with the face position input vector to learn the person-centric heatmap. Similarly to face position, the ground truth used for learning the heatmap is available as a gaze target position in (x, y) coordinates which is quantized into 10 grids in each dimension.

Lastly, the input vectors to the last layer of each pathway are concatenated and go into the final fully connected layer to estimate the "strength" of the fixation ie. how likely it is that the person is actually fixating at a gaze target within the observable scene. The training label for this value is equal to 1 for a fixation inside of the image and 0 when the subject is looking outside of the scene. We also explore alternative model architectures and restrict our training to a subset of the three datasets.

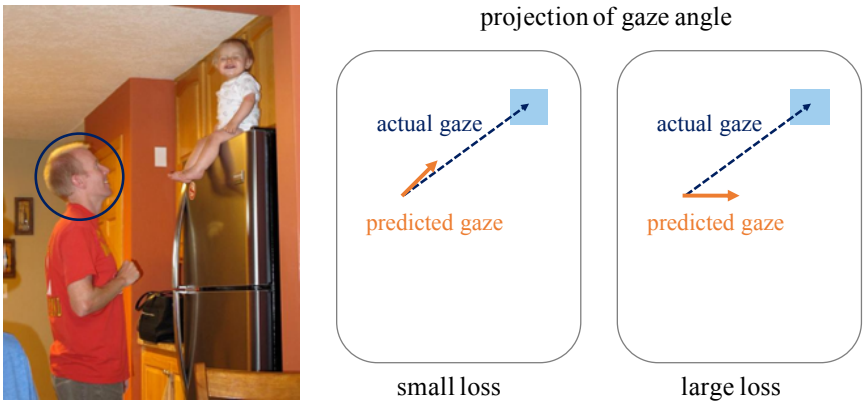


Figure 4.5: **Illustration of the project and compare loss.** If the estimated angle is close to the actual one, the projected gaze angle on the image should also be close to the vector connecting the head position to the gaze target.

Loss

As our model predicts gaze angle, saliency map and the fixation likelihood, we need to apply appropriate loss functions for each task. For the angle regression task we use an *L1 loss*, and for the other two tasks we use a *cross entropy loss*. Moreover, we recognize that the gaze angle and fixation target predictions are closely related. Based on their relationship additional constraints can be imposed to augment the training loss signal. Namely, when the subject is looking at a target, the actual gaze is a ray from the subject's head to the gaze target. This ray can be projected onto the image. It becomes a 2D vector coming from the subject's head to the target exemplified by the blue vector in Figure 4.5. If the estimated angle is close to the actual one, the projected gaze angle on the image (orange vector in Figure 4.5) should also be close to the blue vector. The proximity of the two vectors is measured using the cosine distance. We call this the *project and compare loss*.

Cross-Domain Datasets and Training Procedure

The largest challenge in training our model is the lack of availability of training examples. Although there are a couple of existing datasets that are suitable for training certain parts of our network, no single dataset contains all of the information that we need to train the full

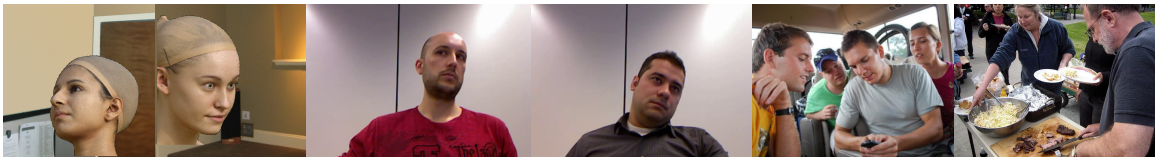


Figure 4.6: **Examples of datasets used to train the model.** Left two: SynHead, middle two: EYEDIAP, right two: GazeFollow.

model. Therefore, we leverage three different datasets, namely, GazeFollow [129], EYEDIAP [159], and SynHead [160]. We selectively train different sub-parts of our network at a time depending on the available supervisory information within a training batch. See Figure 4.6 to see sample images from each dataset.

GazeFollow [129] is a real-world image dataset with manual annotations of the locations where people are looking. The images are taken from other major datasets such as MS COCO [161] and PASCAL [162]. As a result, the images cover a wide range of scenes, people, and gaze directions. However, the actual 3D gaze angles are not available. Furthermore, images where subjects are looking outside of the image frame are not distinguished and all images have a fixation annotation inside of the frame. Although it is mentioned in [129] that if the annotators indicated that the person was looking outside the image, the image would be discarded, we notice that there are a considerable number of images in which persons appear to be looking outside of the frame. Therefore, we added additional annotations to this dataset in the form of a binary indicator label for "looking inside" or "looking outside" for every image. In total, we identified 14,564 images correspond to the "looking outside" case which is approximately 11.6% of the total training samples. We have publicly released our additional annotations along with this paper.

EYEDIAP [159] dataset is designed for the evaluation of the gaze estimation task. It has videos of 16 different subjects with full face and background visible in a laboratory environment. Each subject was asked to look at a specific target point on a monitor screen and the 3D gaze angle was annotated by leveraging camera calibration and face depth measurement from depth camera. This dataset contains precise 3D gaze angles for frames

where the person is fixating the target point. The dataset also contains video of the subjects looking at a 3D ball target instead of 2D screen target point, but we exclude these ball sessions from our experiments in order to conduct a fairer comparison with prior work. We randomly hold out four subjects for test and use the rest of the sessions for training. Since subjects were looking at a screen, all of the frames can be considered as looking outside the image. However, since the dataset has been collected in a controlled setting the backgrounds are primarily white and there is not a lot of variety in lighting or pose. Also, measured gaze angles range between -40° to 40° which is rather limited.

NVIDIA SynHead [160] is a synthetic dataset created for the head pose estimation task. The dataset contains 510,960 frames of 70 head motion tracks rendered using 10 individual head models. The gaze of the head is fixed and aligned with the head pose, thus we use the labeled 3D head pose as the gaze angle ground truth. One of the advantages of a synthetic dataset is the ability to insert different images in the background. We randomly generated 15% from the total frames augmented with provided natural scene backgrounds and regard all as "looking outside" examples. The main reason we include SynHead in training is because it complements the EYEDIAP dataset, as the angle ranges are larger, between -90° and 90°, and it can include more diverse backgrounds. Since head pose estimation is not a focus of this paper we do not set aside a test set and use SynHead entirely for training. Dataset details are also summarized in Table 4.1.

Since each dataset is relevant only to certain subtasks, we only update the relevant parts of the network based on which dataset the training sample is from, while freezing other irrelevant layers during back-propagation. Specifically, when learning gaze angle estimation, we only update the angle pathway (b) and (d) in Figure 4.4, when learning saliency we update the scene pathway (a), (b) and (c) while freezing all other layers. Similarly, when training fixation likelihood we only update the layer (e) in Figure 4.4. We found that this selective back-propagation scheme is critical in achieving good performance.

In every batch, we draw random samples from all of the datasets shuffled together

Table 4.1: Datasets used in the experiments and the number of samples in the training and testing split, as well as the percentage of each split containing people looking inside vs outside.

Dataset	Training set		Test set	
		in vs out		in vs out
GazeFollow [129]	125,557	88.4% vs 11.6%	4,782	100% vs 0%
EYEDIAP [159]	72,613	0% vs 100%	18,153	0% vs 100%
SynHead [160]	75,400	0% vs 100%	-	-
MMDB [11]	-	-	4,965	41.4% vs 58.6%

and perform three separate back-propagation for the three outputs as just described. In the beginning, both convolutional pathways were initialized using a ResNet50 model pre-trained on the ImageNet classification task [163]. We use the Adam optimization algorithm with a learning rate of $2.5e^{-4}$ and a batch size of 36. Training usually converges within 12 epochs.

Table 4.2: **Spatial module evaluation** on the GazeFollow dataset for single image gaze target prediction.

Method	AUC	Avg Distance	Min Distance
Random	0.504	0.484	0.391
Center	0.633	0.313	0.230
Judd [164]	0.711	0.337	0.250
GazeFollow [129]	0.878	0.190	0.113
Chong [130]	0.896	0.187	0.112
Ours	**0.920**	**0.143**	**0.079**
Human	0.924	0.096	0.040

Table 4.3: **Quantitative model evaluation** on our VideoAttentionTarget dataset.

Method	*spatial*		*out of frame*
	AUC	L^2 Dist.	Average Precision
Random	0.505	0.458	0.621
Fixed bias	0.728	0.326	0.624
Chong [130]	0.850	0.193	0.705
Chong [130]+LSTM	0.863	0.171	0.712
No head position	0.835	0.169	0.827
No head features	0.758	0.258	0.714
No attention map	0.717	0.226	0.774
No fusion	0.853	0.165	0.817
No temporal	0.864	0.147	0.858
Ours full	**0.882**	**0.128**	**0.860**
Human	0.921	0.051	0.925

4.4 Results

We conducted four experiments to evaluate the performance of our method. Sec. 4.4.1 uses just the spatial component of our model on the GazeFollow dataset. Sec. 4.4.2 uses the full spatiotemporal model on the VideoAttentionTarget dataset. Sec. 4.4.3 uses the model output to classify clinically-relevant social behaviors in a sample of toddlers. Sec. 4.4.4 uses the model to detect shared attention in the VideoCoAtt dataset. *Our method produces state-of-the-art results on all datasets in all experiments.*

Evaluation in Secs. 4.4.1 and 4.4.2 follow the protocol from [130] which we review.

There are three performance meausures: AUC, Distance, and Out-of-Frame AP. **AUC:** Each cell in the spatially-discretized image is classified as gaze target or not. The ground truth comes from thresholding a Gaussian confidence mask centered at the human annotator's target location. The final heatmap provides the confidence score which is evaluated at different thresholds in the ROC curve. **Distance:** L^2 distance between the annotated target location and the prediction given by the pixel of maximum value in the heatmap, with image width and height normalized to 1. AUC and Distance are computed whenever there is an in-frame ground truth gaze target (the heatmap always has a max). **Out-of-Frame AP:** The average precision is computed for the prediction score from the "out-of-frame" scalar (described in Sec. 4.3.2) against the ground truth, computed in every frame. Final measures are the average across the annotators. We also evaluate the performance of the annotators (**Human** performance) across all three measures. This is done by comparing annotator predictions in all pairs and averaging them. This is analogous to the kappa score used to measure inter-rater reliability, but specialized for our performance measures.

4.4.1 Spatial Module Evaluation

We evaluate the static part of our model on single image gaze target prediction using the GazeFollow dataset [129], and compare against prior methods. GazeFollow contains annotations of person heads and gaze locations in a diverse set of images. To train the model, we used the annotation labels from [130] which were adjusted to additionally specify whether the annotated gaze target is out-of-frame. In order to make a fair comparison, we only use the GazeFollow dataset for training and do not include our new dataset.

The results in Table 4.2 demonstrate the value of our architectural choices in the spatial model component. We outperform previous methods by a significant margin. In fact, our AUC of 0.92 is quite close to the AUC of 0.924 obtained by human. Qualitatively, visualization of the learned weights of the attention layer reveals that the model has learned to effectively make use of the facial orientation information for weighting scene features,

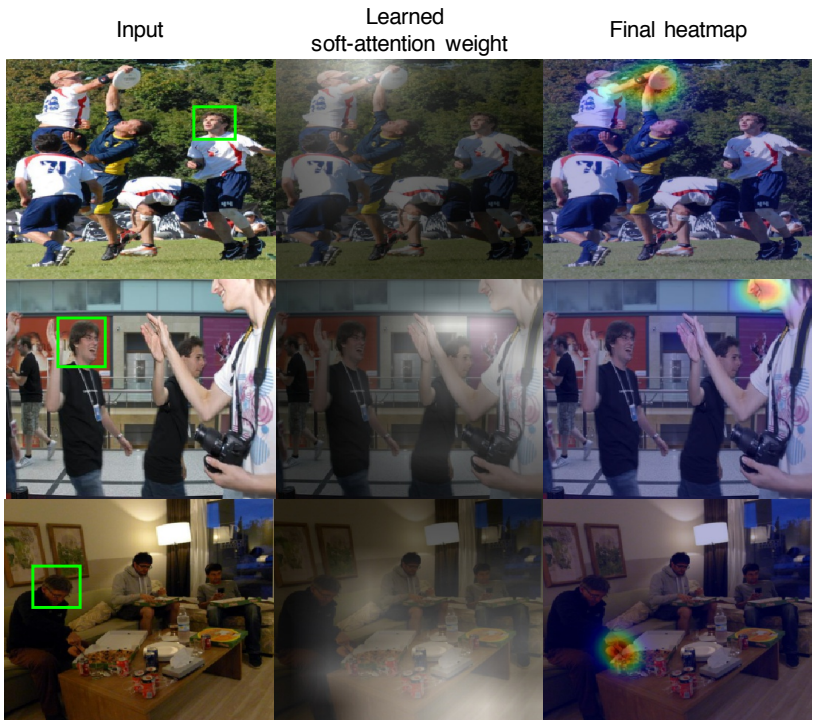


Figure 4.7: **Visualization of head-conditioned attention** with corresponding input and final output. The attention layer captures and leverages the head pose information to regulate the model's prediction.

as shown in Fig. 4.7.

4.4.2 Spatiotemporal Model Evaluation

We evaluate our full spatio-temporal model on the new *VideoAttentionTarget* dataset. Table 4.3 summarizes the experimental results. The first block of rows shows baseline tests and comparison with previous methods; *Random* is when the prediction is made at 50% chance, and *Fixed bias* is when the bias present in the dataset (Fig 4.2b) is utilized. The method of [130], which is the existing non-temporal gaze target estimator, is compared both as-is and using an additional LSTM layer on top. The second set of rows in the table shows ablation study results by disabling key components of our model one at a time; *No head position* is when the head position image is not used. *No head features* is when the head feature map from the Head Conv module is not provided. In this case, the attention map is made using only the head position. *No attention map* is when attention map is not

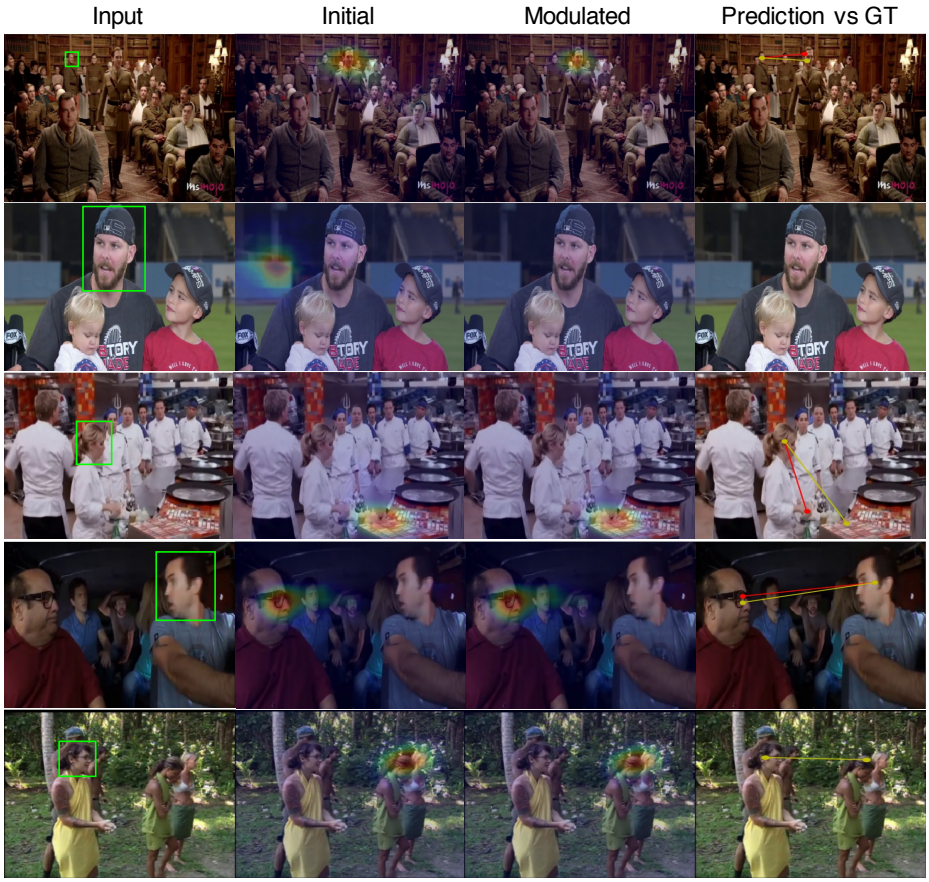


Figure 4.8: **Gaze target prediction results on example frames.** *Initial* denotes the first output of the deconvolution, *Modulated* shows the adjusted heatmap after modulation. Final prediction (yellow) and ground truth (red) are presented in the last column. Rows 1, 3, 4 depict properly predicted within-image gaze target, row 2 shows correctly identified nonexistent gaze target in frame, and the last row is an example of failure case where it predicts a fixated target in the image due to the lack of sense of depth when the subject is actually looking outside.

produced therefore the scene feature map is uniformly weighted. *No fusion* is when the head feature map is only used to produce attention map and not concatenated with scene feature map for encoding. *No temporal* is when ConvLSTM is not used. This quantitative analysis demonstrates that our proposed model strongly outperforms previous methods as well as the presented baselines. All components of the model are crucial to achieving the best performance, and the head convolutional pathway and the attention mechanism were found to have the biggest contribution. Qualitative results are presented in Fig. 4.8.

4.4.3 Detecting the Social Bids of Toddlers

Motivation Eye contact and joint attention are among the earliest social skills to emerge in human development [165], and are closely-related to language learning [166] and socio-emotional development [167]. Children with autism exhibit difficulty in modulating gaze during social interactions [22, 168, 169], and social gaze is assessed as part of the diagnosis and treatment of autism. This is usually done qualitatively or through laborious manual methods. The ability to automatically quantify children's social gaze would offer significant benefits for clinicians and researchers.

Toddler Dataset We collected a dataset of 20 toddlers (10 with an autism diagnosis, 10 female, mean age 36.4 months) who were video-recorded during dyadic social interactions. Each participant completed an assessment known as the ESCS [12], which was administered by trained examiners. The ESCS is a semi-structured play protocol designed to elicit nonverbal social communication behaviors. Parents provided written consent for data collection and processing.

Five expert raters annotated all of the child's gaze shifts consisting of looks to toys and looks to the examiner's face at the frame level. Based on this per-frame annotation, a toy-to-eyes gaze shift event is inferred if the gaze target changes from the toy to the examiner's face within 700 milliseconds. In total, the dataset contains 623 shift events during 221 minute long recording. Our task was to detect these toy-to-eyes gaze shifts and reject all other types of gaze shifts which the child made during the session. The toy-to-eyes shifts are relevant to child development because they can be further classified into different types of joint attention based on the context in which they are produced [12]. Joint attention is a key construct for the development of social communication. Our experiment provides preliminary evidence for the feasibility of automatically identifying such gaze-based joint attention events from video.

Experimental setup and results Given the toddlers dataset, we conducted experiments to see how an automated method can be used to retrieve gaze shift events. Two types of

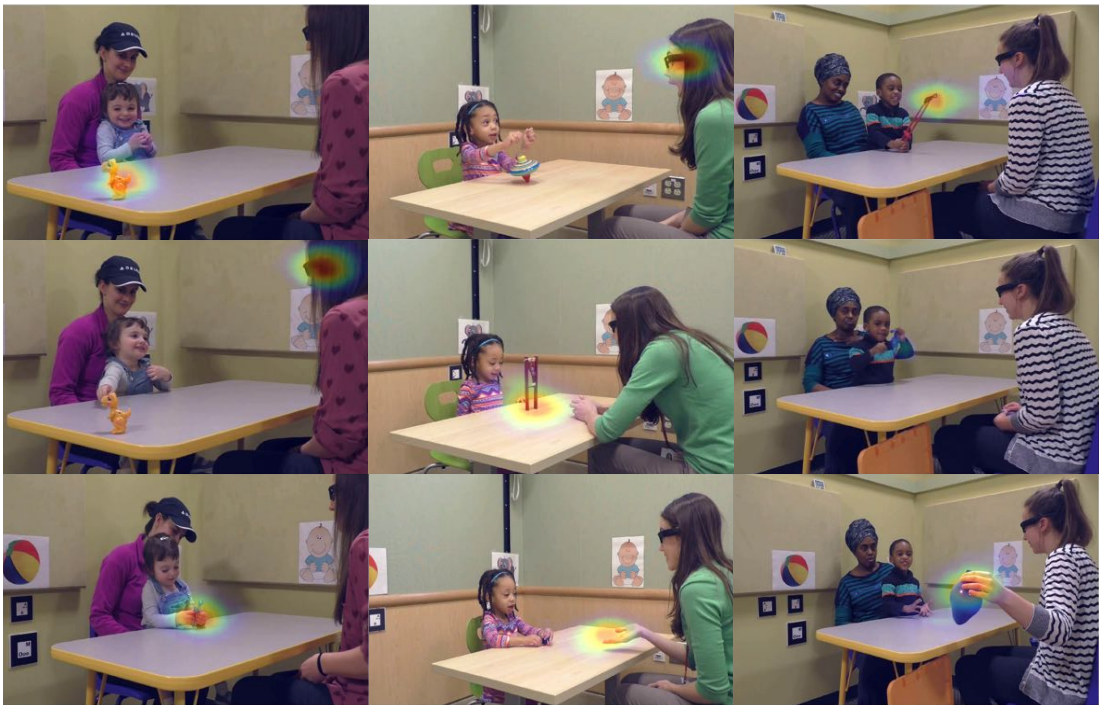


Figure 4.9: **Heatmap output** of our model on toddlers video.

approaches for shift detection are explored. The first approach is to detect a shift in a two-step process where we initially classify the type of attended object - among toy, eyes, and elsewhere - in every frame with a ResNet-50 image classifier, and then apply an event classifier on top of it over a temporal window to conclusively find the gaze shift from toy to eyes. A random forest model is used for the event classifier. For the second approach, we try detecting a shift event in an end-to-end manner, using the I3D model [170] since gaze shift can be viewed as a special case of a human action.

For both approaches, we compare shift detection performance when the inputs to the models are 1. the plain RGB image, 2. image and head position, and 3. image and heatmap produced by our attention network (Fig. 4.9). For 2 and 3 the head position or heatmap is concatenated depth-wise to the RGB image as a 4th channel in grayscale. CNN layers of ResNet were pretrained on ImageNet [163] and those of I3D were pretrained on Kinetics [10]. The child's head was detected and recognized using [99]. A sliding window size of 64 frames was used during training. For validation, we adopted 5-fold subject-wise cross

validation in which 4 subjects were held out in each validation set.

Table 4.4: **Gaze coding detection results** on the toddlers dataset. As shown, our heatmap feature indeed improves shift detection when used along with image in a standard classification paradigm.

Method	Detection Approach	Prec.	Rec.
Random		0.034	0.503
ResNet on RGB	image feature & classifier	0.541	0.567
ResNet on RGB+head		0.598	0.575
ResNet on RGB**+heatmap**		**0.708**	**0.759**
I3D on RGB	end-to-end	0.433	0.506
I3D on RGB+head		0.475	0.500
I3D on RGB**+heatmap**		0.559	0.710
Human (clinical experts)		0.903	0.922

Table 4.4 summarizes the results of our experiment with the precision and recall of gaze shift detection. Interestingly, in general the 2D-CNN-based approach outperformed the 3D-CNN method, which is presumably due to the complexity of I3D and relatively less training data. Although as a human we can easily tell a child's gaze behavior given the plain image, our experiments show that the added heatmap information is a useful feature to have for this task. Nevertheless, there still exists a noticeable gap between human performance, implying the need for further research on this problem. The next possible steps include the modeling of depth and body posture and addition of different view points.

4.4.4 Detecting Shared Attention in Social Scenes

As an additional application of our system on real-world problems, we apply our model to infer shared attention in the scene. We use the VideoCoAtt dataset [138] to benchmark our performance on this task. This dataset has 113,810 test frames that are annotated with the target location when it is simultaneously attended by two or more people.

Given that our model does not have a head detection module as in [138], we use a publicly available off-the-shelf head detector [137] to automatically generate the input subject

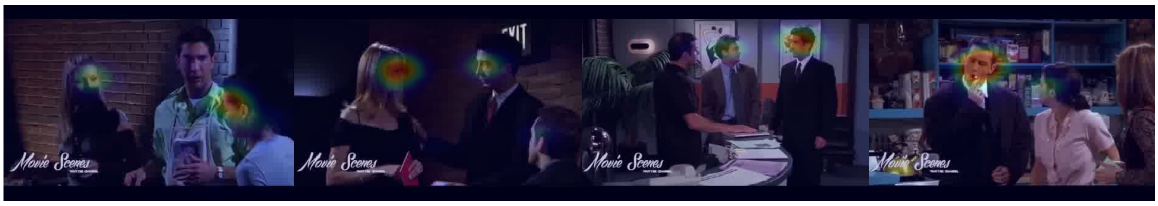


Figure 4.10: **Constructed shared attention map** by adding up individual heatmaps of all people in the image. Samples are from VideoCoAtt dataset.

head positions. We did not choose to fine-tune our model on the VideoCoAtt dataset since their annotations do not naturally translate to a dense single-subject-target annotations that our model can be trained with.

Our method is evaluated on the following two tasks: 1. location prediction (spatial) and 2. interval detection (temporal) of shared attention. For the localization task, we first add up individual heatmaps of all people in the frame and aggregate them into a single shared attention confidence map (examples in Fig. 4.10). Then, the L^2 distance is computed between the pixel position of the maximum and the center of the ground truth target. For the interval detection task, we regard the frame as presenting a shared attention case if the maximum pixel's score is above certain threshold. Specifically, we choose a threshold of 1.8, because a single heatmap from our model can have a maximum value of 1 at the fixated location and when another heatmap is added to the same location its value becomes 2. The threshold value of 1.8 is chosen to make a room for slight misalignments between multiple fixations.

As a result, our method achieves state-of-the-art results on both tasks, as shown in Table 4.5. This outcome is surprising since the model by Fan et al. [138] was formulated specifically to detect shared attention whereas ours was not. However, it must also be noted that there exist differences in experimental setup such as the head detector and the training data, thus it should not be regarded as a direct comparison between the two models. Here, we intend to demonstrate the potential use of our model for recognizing higher-level social gaze, and it is encouraging that we can achieve great performance without tweaking the model for this specific problem.

Table 4.5: **Shared attention detection results** on the VideoCoAtt dataset. The interval detection task is evaluated with prediction accuracy and the localization task is measured with L^2.

Method	Pred. Acc. (%)	L^2 Distance
Random	50.8	286
Fixed bias	52.4	122
GazeFollow [129]	58.7	102
Gaze+Saliency [171]	59.4	83
Gaze+Saliency [171]+LSTM	66.2	71
Fan [138]	71.4	62
Ours	**83.3**	**57**

4.5 Conclusion

We have presented a new deep architecture and a novel *VideoAttentionTarget* dataset for the task of detecting the time-varying attention targets for each person in a video. Our model is designed to allow the face to direct the learning of gaze-relevant scene regions, and our new dataset makes it possible to learn the temporal evolution of these features. The strong performance of our method on multiple benchmark datasets and a novel social gaze recognition task validates its potential as a useful tool for understanding gaze behavior in naturalistic human interactions.

CHAPTER 5
CONCLUSION

This thesis investigated different computational ways of measuring human visual attention during face-to-face interactions from video and how it can be used to identify socially meaningful gaze behaviors such as eye contact and joint attention.

Specifically, three methods were presented. First, I presented methods for the detection of looks to camera from an egocentric view of the social partner and its use for eye contact detection. Second, I presented a 3D head tracking method and its application for measuring attentional shifts. Third, I presented spatiotemporal deep neural network that learns time-varying attention targets from video and its utility for social gaze behavior recognition tasks in 3rd-person videos.

Through these works, I have demonstrated that it is possible to measure the moments of eye contact at the human expert level, and it is feasible to detect gaze targets from a diverse set of social interaction videos. Its significance comes from its ability to automatically quantify behaviors from videos of face-to-face interactions, which greatly benefits the scalability of the system, removing the burden of using wearable eye trackers. Experimental results have been extensively validated on a large-scale dataset of social interactions in clinical assessment settings to demonstrate its usability for clinical and academic purposes.

While these results are promising, there remain a number of technical limitations which need to be overcome in the future research efforts. First of all, there is a fundamental relationship between image resolution and depth since the farther away the person is from the camera, the lower the image quality will be around the face and the eyes, making it harder to disambiguate his/her gaze direction. This could particularly be a challenge for home deployment as people will be freely moving around and therefore the relative person-camera distance may not be secured, along with the other potential issues caused by the

limited field-of-view, occlusion, and camera motion. Advancements on portable camera technology such as the development of wide-angle, high-resolution imager and long-life batteries could greatly benefit our application in this aspect.

From the algorithmic point of view, there are two major challenges. First one is regarding the stability of attention target prediction. Current attention model struggles with the failures that arise from the difficulty of distinguishing the foreground from the background. For example, when there is a salient picture in the background that coincides with the direction of sight of a forward-looking subject, the picture behind the subject can be classified as the gaze target by mistake. Another challenging case is when the scene is cluttered with a lot of objects, in which case the system makes confusion among many attention target candidates. Along this line, it is generally observed that the human face is chosen as the answer when there are many competing targets nearby, which is primarily due to the bias in training dataset, which might not always be true. To address these problems we need to develop machine learning solutions that can reason about scene layout and contextual cues, and better handle data bias. Second challenge is related to the need for manual intervention. Current major barrier for the system to being completely automatic is the fact that manual interference is still required during the process of maintaining the track of one subject among all people in the video. Modern multi-target tracking or person re-identification methods can help resolving some of the issues if the camera is static, but it still remains a great challenge for the egocentric view because it is difficult to automatically associate detections across far-apart frames in limited, fast transitioning, skewed viewpoint. A novel automated approach needs to be established to effectively associate entities in dynamic views to resolve this issue.

My thesis thus far has focused on the visual attention modality. However, human social intelligence is a much more complex construct, involving various modalities which are not only gaze, but also gestures, body pose, facial expression, speech, etc, and also how different modalities are coordinated in time and space. Towards building general computational

social intelligence, we have a lot of research questions to investigate such as; What are the available datasets and methods that we can leverage to solve this problem? How can we create a new public dataset specific in social dimension that can contribute to research community and drive innovations? What are the necessary sensors and setups that best capture natural social interactions? What kind of computational models can we build to measure multi-modal social behavior? To what extent can one modality help modeling other modalities? What are the underlying variables that govern common aspects of behavior and how can we discover them from data?

Outside the experimental settings mainly considered in the thesis, there are many other exciting areas that could potentially benefit from the presented approaches, with broad implications on different research topics and opportunities for new applications. Examples include the automatic analysis of turn-taking and social roles in group interactions, skill assessment in professional settings such as job interviews, and the development of social robots that are able to interact and collaborate naturally with humans using nonverbal social signals.

REFERENCES

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition.," in *British Machine Vision Conference*, vol. 1, 2015, p. 6.

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, Springer, 2016, pp. 499–515.

[3] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[6] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[7] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[9] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*, Springer, 2016, pp. 510–526.

[10] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *ArXiv preprint arXiv:1705.06950*, 2017.

[11] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O Ousley, Y. Li, C. Kim, *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3414–3421.

[12] P. Mundy, C. Delgado, J. Block, M. Venezia, A. Hogan, and J. Seibert, "Early social communication scales (escs)," *Coral Gables, FL: University of Miami*, 2003.

[13] C. Lord, P. C. DiLavore, and K. Gotham, *Autism diagnostic observation schedule*. Western Psychological Services Torrance, CA, 2012.

[14] R. Grzadzinski, T. Carr, C. Colombi, K. McGuire, S. Dufek, A. Pickles, and C. Lord, "Measuring changes in social communication behaviors: Preliminary development of the brief observation of social communication change (boscc)," *Journal of Autism and Developmental Disorders*, vol. 46, no. 7, pp. 2464–2479, 2016.

[15] K. Kaye and A. Fogel, "The temporal structure of face-to-face communication between mothers and infants.," *Developmental psychology*, vol. 16, no. 5, p. 454, 1980.

[16] S. P. Vecera and M. H. Johnson, "Gaze detection and the cortical processing of faces: Evidence from infants and adults," *Visual cognition*, vol. 2, no. 1, pp. 59–87, 1995.

[17] T. Farroni, M. H. Johnson, E. Menon, L. Zulian, D. Faraguna, and G. Csibra, "Newborns' preference for face-relevant stimuli: Effects of contrast polarity," *Proceedings of the National Academy of Sciences*, vol. 102, no. 47, pp. 17 245–17 250, 2005.

[18] V. M. Reid, K. Dunn, R. J. Young, J. Amu, T. Donovan, and N. Reissland, "The human fetus preferentially engages with face-like visual stimuli," *Current Biology*, vol. 27, no. 12, pp. 1825–1828, 2017.

[19] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, pp. 289–304, 1965.

[20] C. L. Kleinke, "Gaze and eye contact: A research review.," *Psychological Bulletin*, vol. 100, no. 1, p. 78, 1986.

[21] P. Mundy and L. Newell, "Attention, joint attention, and social cognition," *Current directions in psychological science*, vol. 16, no. 5, pp. 269–274, 2007.

[22] P. Mundy, M. Sigman, J. Ungerer, and T. Sherman, "Defining the social deficits of autism: The contribution of non-verbal communication measures," *Journal of child psychology and psychiatry*, vol. 27, no. 5, pp. 657–669, 1986.

[23] R. J. Hagerman, K. Amiri, and A. Cronister, "Fragile x checklist," *American journal of medical genetics*, vol. 38, no. 2-3, pp. 283–287, 1991.

[24] S. R. Miller, C. J. Miller, J. S. Bloom, G. W. Hynd, and J. G. Craggs, "Right hemisphere brain morphology, attention-deficit hyperactivity disorder (adhd) subtype, and social comprehension," *Journal of child neurology*, vol. 21, no. 2, pp. 139–144, 2006.

[25] D. M. Riby and P. J. Hancock, "Viewing it differently: Social scene perception in williams syndrome and autism," *Neuropsychologia*, vol. 46, no. 11, pp. 2855–2860, 2008.

[26] X. Fu, E. E. Nelson, M. Borge, K. A. Buss, and K. Pérez-Edgar, "Stationary and ambulatory attention patterns are differentially associated with early temperamental risk for socioemotional problems: Preliminary evidence from a multimodal eye-tracking investigation," *Development and Psychopathology*, pp. 1–18, 2019.

[27] L. Ezpeleta, R. Granero, N. de la Osa, and J. M. Domènech, "Clinical characteristics of preschool children with oppositional defiant disorder and callous-unemotional traits," *PloS one*, vol. 10, no. 9, e0139346, 2015.

[28] M. Rutter, A Le Couteur, and C Lord, "Autism diagnostic interview-revised," *Los Angeles, CA: Western Psychological Services*, vol. 29, p. 30, 2003.

[29] J. Holler and K. H. Kendrick, "Unaddressed participants' gaze in multi-person interaction: optimizing recipiency," *Frontiers in Psychology*, vol. 6, no. Feb, pp. 1–14, 2015.

[30] S. Ho, T. Foulsham, and A. Kingstone, "Speaking and listening with the eyes: gaze signaling during dyadic interactions," *PLoS ONE*, vol. 10, no. 8, pp. 1–18, 2015.

[31] S. L. Rogers, C. P. Speelman, O. Guidetti, and M. Longmuir, "Using dual eye tracking to uncover personal gaze patterns during social interaction," *Scientific Reports*, no. February, pp. 1–9, 2018.

[32] J. M. Franchak, K. S. Kretch, K. C. Soska, and K. E. Adolph, "Head-mounted eye-tracking: A new method to describe infant looking," *Child Development*, vol. 82, no. 6, pp. 1738–1750, 2011.

[33] C. Yu and L. B. Smith, "Hand-eye coordination predicts joint attention," *Child Development*, vol. 88, no. 6, pp. 2060–2078, 2017.

[34] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg, "Detecting eye contact using wearable eye-tracking glasses," in *Proceedings of the ACM Conference on Ubiquitous Computing*, ACM, 2012, pp. 699–704.

[35] Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg, "Detecting bids for eye contact using a wearable camera," in *Proc. 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, 2015.

[36] S. R. Edmunds, A. Rozga, Y. Li, E. A. Karp, L. V. Ibanez, J. M. Rehg, and W. L. Stone, "Brief report: Using a point-of-view camera to measure eye gaze in young children with autism spectrum disorder during naturalistic social interactions: A pilot study," *Journal of Autism and Developmental Disorders*, pp. 1–7, 2017.

[37] R. M. Jones, A. Southerland, A. Hamo, C. Carberry, C. Bridges, S. Nay, E. Stubbs, E. Komarow, C. Washington, J. M. Rehg, *et al.*, "Increased eye contact during conversation compared to play in children with autism," *Journal of autism and developmental disorders*, vol. 47, no. 3, pp. 607–614, 2017.

[38] E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R. M. Jones, A. Rozga, and J. M. Rehg, "Detecting gaze towards eyes in natural social interactions and its use in child assessment," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 43, 2017.

[39] Y. Mitsuzumi, A. Nakazawa, and T. Nishida, "Deep eye contact detector: robust eye contact bid detection using convolutional neural network," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017, pp. 1–12.

[40] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.

[41] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, 2018.

[42] M. Wang and W. Deng, "Deep face recognition: A survey," *ArXiv preprint arXiv:1804.06655*, 2018.

[43] Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg, "Detecting bids for eye contact using a wearable camera," in *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, vol. 1, 2015, pp. 1–8.

[44] J. S. Shell, R. Vertegaal, D. Cheng, A. W. Skaburskis, C. Sohn, A. J. Stewart, O. Aoudeh, and C. Dickie, "Ecsglasses and eyepliances: Using attention to open sociable windows of interaction," in *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications (ETRA)*, ACM, 2004, pp. 93–100.

[45] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," in *Proceedings of the 26th Annual*

ACM Symposium on User Interface Software and Technology (UIST), ACM, 2013, pp. 271–280.

[46] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[47] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," *ArXiv preprint arXiv:1611.08860*, 2016.

[48] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176–2184.

[49] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1821–1828.

[50] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4511–4520.

[51] M. Chita-Tegmark, "Social attention in asd: A review and meta-analysis of eye-tracking studies," *Research in Developmental Disabilities*, vol. 48, pp. 79–93, 2016.

[52] K. Pierce, D. Conant, R. Hazin, R. Stoner, and J. Desmond, "Preference for geometric patterns early in life as a risk factor for autism," *Archives of General Psychiatry*, vol. 68, no. 1, pp. 101–109, 2011.

[53] M. Hosozawa, K. Tanaka, T. Shimizu, T. Nakano, and S. Kitazawa, "How children with specific language impairment view social situations: An eye tracking study," *Pediatrics*, vol. 129, no. 6, e1453–e1460, 2012.

[54] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Archives of General Psychiatry*, vol. 59, no. 9, pp. 809–816, 2002.

[55] K. Chawarska and F. Shic, "Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder," *Journal of Autism and Developmental Disorders*, vol. 39, no. 12, p. 1663, 2009.

[56] Q. Guillon, N. Hadjikhani, S. Baduel, and B. Rogé, "Visual social attention in autism spectrum disorder: Insights from eye tracking studies," *Neuroscience & Biobehavioral Reviews*, vol. 42, pp. 279–297, 2014.

[57] B. Noris, J. Nadel, M. Barker, N. Hadjikhani, and A. Billard, "Investigating gaze of children with asd in naturalistic settings," *PloS One*, vol. 7, no. 9, e44144, 2012.

[58] S. Magrelli, P. Jermann, N. Basilio, F. Ansermet, F. Hentsch, J. Nadel, and A. Billard, "Social orienting of children with autism to facial expressions and speech: A study with a wearable eye-tracker in naturalistic settings," *Frontiers in Psychology*, vol. 4, p. 840, 2013.

[59] N. J. Sasson and J. T. Elison, "Eye tracking young children with autism," *Journal of Visualized Experiments*, no. 61, e3675–e3675, 2012.

[60] L. Zwaigenbaum, S. Bryson, and N. Garon, "Early identification of autism spectrum disorders," *Behavioral Brain Research*, vol. 251, pp. 133–146, 2013.

[61] A. M. Wetherby and B. M. Prizant, *Communication and symbolic behavior scales: Developmental profile*. Paul H Brookes Publishing, 2002.

[62] D. L. Robins, D. Fein, M. L. Barton, and J. A. Green, "The modified checklist for autism in toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 31, no. 2, pp. 131–144, 2001.

[63] *Mangold interact*, Accessed: 2019-09-03.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[65] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 162–175, 2019.

[66] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ACM, 2014, pp. 255–258.

[67] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1548–1557.

[68] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[69] D. J. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *Journal of pharmacokinetics and biopharmaceutics*, vol. 15, no. 6, pp. 657–680, 1987.

[70] A. Rozga, A. Southerland, M. McCall, E. Stubbs, M. Silverman, E. Ajodan, K. Chanda, E. Chong, J. M. Rehg, and R. M. Jones, "Characterizing temporal-contextual effects on social and object-directed attention in asd via high-density video coding," *International society for autism research (INSAR)*, 2018.

[71] V. Hus, K. Gotham, and C. Lord, "Standardizing ados domain scores: Separating severity of social affect and restricted and repetitive behaviors," *Journal of autism and developmental disorders*, vol. 44, no. 10, pp. 2400–2412, 2014.

[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[73] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Continuous conditional neural fields for structured regression," in *European Conference on Computer Vision*, Springer, 2014, pp. 593–608.

[74] L. S. Nguyen and D. Gatica-Perez, "I would hire you in a minute," in *Proceedings ACM International Conference on Multimodal Interaction (ICMI 15)*, 2015, pp. 51–58.

[75] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: the role of what you say and how you say it," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, 2015.

[76] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[77] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[78] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[79] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided

detection of mammographic lesions," *Medical image analysis*, vol. 35, pp. 303–312, 2017.

[80] R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, D. Hanel, M. Gardner, A. Gupta, R. Hotchkiss, *et al.*, "Deep neural network improves fracture detection by clinicians," *Proceedings of the National Academy of Sciences*, vol. 115, no. 45, pp. 11 591–11 596, 2018.

[81] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, p. 65, 2019.

[82] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard, "Mood meter: Counting smiles in the wild," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM, 2012, pp. 301–310.

[83] J. Hashemi, M. Tepper, T. Vallin Spina, A. Esler, V. Morellas, N. Papanikolopoulos, H. Egger, G. Dawson, and G. Sapiro, "Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants.," *Autism research and treatment*, vol. 2014, 2014.

[84] Z. Hammal, J. F. Cohn, and D. S. Messinger, "Head movement dynamics during play and perturbed mother-infant interaction," *IEEE transactions on affective computing*, vol. 6, no. 4, pp. 361–370, 2015.

[85] J. Hashemi, G. Dawson, K. L. Carpenter, K. Campbell, Q. Qiu, S. Espinosa, S. Marsan, J. P. Baker, H. L. Egger, and G. Sapiro, "Computer vision analysis for quantification of autism risk behaviors," *IEEE Transactions on Affective Computing*, 2018.

[86] K. Campbell, K. L. Carpenter, J. Hashemi, S. Espinosa, S. Marsan, J. S. Borg, Z. Chang, Q. Qiu, S. Vermeer, E. Adler, *et al.*, "Computer vision analysis captures atypical attention in toddlers with autism," *Autism*, vol. 23, no. 3, pp. 619–628, 2019.

[87] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu, "3d human sensing, action and emotion recognition in robot assisted therapy of children with autism," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2158–2167.

[88] W. Zhu and H. Deng, "Monocular free-head 3d gaze tracking with deep learning and geometry constraints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3143–3152.

[89] G. Liu, Y. Yu, and J.-M. Odobez, "A differential approach for gaze estimation with calibration.," in *British Machine Vision Conference*, 2018.

[90] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: A review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[91] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social robotics," in *Springer handbook of robotics*, Springer, 2016, pp. 1935–1972.

[92] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin, "Behavioral imaging and autism," *IEEE Pervasive Computing*, vol. 13, no. 2, pp. 84–87, 2014.

[93] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[94] C. Breazeal, "Role of expressive behaviour for robots that learn from people," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1535, pp. 3527–3538, 2009.

[95] J. F. Ferreira and J. Dias, "Attentional mechanisms for socially interactive robots–a survey," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 110–125, 2014.

[96] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, IEEE, 2011, pp. 30–35.

[97] D. S. Messinger, M. H. Mahoor, S.-M. Chow, and J. F. Cohn, "Automated measurement of facial expression in infant–mother interaction: A pilot study," *Infancy*, vol. 14, no. 3, pp. 285–305, 2009.

[98] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, IEEE, vol. 1, 2015, pp. 1–8.

[99] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[100] J.-S. Jang and T. Kanade, "Robust 3d head tracking by online feature registration," in *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2008.

[101] S. Choi and D. Kim, "Robust head tracking using 3d ellipsoidal head model in particle filter," *Pattern Recognition*, vol. 41, no. 9, pp. 2901–2915, 2008.

[102] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[103] S. O. Ba and J.-M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 101–116, 2011.

[104] L. Dong, H. Di, L. Tao, G. Xu, and P. Oliver, "Visual focus of attention recognition in the ambient kitchen," in *Asian Conference on Computer Vision*, Springer, 2009, pp. 548–559.

[105] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342.

[106] E. De Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel, "Marker-less deformable mesh tracking for human shape and motion capture," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, IEEE, 2007, pp. 1–8.

[107] J.-G. Wang and E. Sung, "Em enhancement of 3d head pose estimated by point at infinity," *Image and Vision Computing*, vol. 25, no. 12, pp. 1864–1874, 2007.

[108] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[109] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[110] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, IEEE, vol. 1, 2015, pp. 1–8.

[111] S. Tulyakov and N. Sebe, "Regressing a 3d face shape from a single image," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2015, pp. 3748–3755.

[112] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, IEEE, vol. 5, 2004, pp. V–881.

[113] J.-M. Odobez and S. O. Ba, "A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose," in *International Conference on Multi-Media & Expo (ICME07)*, 2007.

[114] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From gaze to focus of attention," in *International Conference on Advances in Visual Information Systems*, Springer, 1999, pp. 765–772.

[115] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 2011, pp. 3457–3464.

[116] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of f-formations," in *Proc. BMVC*, 2011.

[117] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1212–1229, 2008.

[118] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, 2012, pp. 1226–1233.

[119] H. Soo Park and J. Shi, "Social saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4777–4785.

[120] N. J. Emery, "The eyes have it: The neuroethology, function and evolution of social gaze," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[121] S. S. Rajagopalan and R. Goecke, "Detecting self-stimulatory behaviours for autism diagnosis," in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2014, pp. 1470–1474.

[122] P. Wang, G. D. Abowd, and J. M. Rehg, "Quasi-periodic event analysis for social game retrieval," in *2009 IEEE 12th International Conference on Computer Vision*, IEEE, 2009, pp. 112–119.

[123] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[124] R. Mur-Artal, J. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[125] D. F. Abawi, J. Bienwald, and R. Dorner, "Accuracy in optical tracking with fiducial markers: An accuracy function for artoolkit," in *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, 2004, pp. 260–261.

[126] Y. Konishi, Y. Hanzawa, M. Kawade, and M. Hashimoto, "Fast 6d pose estimation from a monocular image using hierarchical pose trees," in *European Conference on Computer Vision*, Springer, 2016, pp. 398–413.

[127] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *Advances in Neural Information Processing Systems*, 2012, pp. 431–439.

[128] M. Land and B. Tatler, *Looking and acting: Vision and eye movements in natural behaviour*. Oxford University Press, 2009.

[129] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" In *Advances in Neural Information Processing Systems*, 2015, pp. 199–207.

[130] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 383–398.

[131] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba, "Following gaze in video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1435–1443.

[132] A. Saran, S. Majumdar, E. S. Shor, A. Thomaz, and S. Niekum, "Human gaze following for human-robot interaction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 8615–8621.

[133] J. Guan, L. Yin, J. Sun, S. Qi, X. Wang, and Q. Liao, "Enhanced gaze following via object detection and human pose estimation," in *26th International Conference on Multimedia Modeling*, 2019.

[134] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax, "Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment," in *ETRA*, 2004.

[135] M. J. Marín-Jiménez, A. Zisserman, M. Eichner, and V. Ferrari, "Detecting people looking at each other in videos," *International Journal of Computer Vision*, vol. 106, no. 3, pp. 282–296, 2014.

[136] C. Palmero, E. A. van Dam, S. Escalera, M. Kelia, G. F. Lichtert, L. P. Noldus, A. J. Spink, and A. van Wieringen, "Automatic mutual gaze detection in face-to-face dyadic interaction videos," *Measuring Behavior 2018*, 2018.

[137] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, "Laeo-net: Revisiting people looking at each other in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3477–3485.

[138] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, "Inferring shared attention in social scene videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6460–6468.

[139] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding human gaze communication by spatio-temporal graph reasoning," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[140] S. E. Bryson, L. Zwaigenbaum, C. McDermott, V. Rombough, and J. Brian, "The autism observation scale for infants: Scale development and reliability data," *Journal of autism and developmental disorders*, vol. 38, no. 4, pp. 731–738, 2008.

[141] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[142] S. O. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 16–33, 2008.

[143] B. Massé, S. Ba, and R. Horaud, "Tracking gaze and visual focus of attention of people involved in social interaction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2711–2724, 2017.

[144] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and why are they looking? jointly inferring human attention and intentions in complex tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6801–6809.

[145] E. Brau, J. Guan, T. Jeffries, and K. Barnard, "Multiple-gaze geometry: Inferring novel 3d locations from gazes observed in monocular video," in *The European Conference on Computer Vision (ECCV)*, 2018.

[146] B. Massé, S. Lathuilière, P. Mesejo, and R. Horaud, "Extended gaze following: Detecting objects in videos beyond the camera field of view," in *14th IEEE Inter-*

national Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019, 2019.

[147] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *Advances in Neural Information Processing Systems*, vol. 1, 2012, pp. 422–430.

[148] J. Hashemi, M. Tepper, T. Vallin Spina, A. Esler, V. Morellas, N. Papanikolopoulos, H. Egger, G. Dawson, and G. Sapiro, "Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants," *Autism research and treatment*, vol. 2014, 2014.

[149] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2074–2083.

[150] E. Chong, A. Southerland, A. Kundu, R. M. Jones, A. Rozga, and J. M. Rehg, "Visual 3d tracking of child-adult social interactions," in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, IEEE, 2017, pp. 399–406.

[151] G. Pusiol, L. Soriano, L. Fei-Fei, and M. C. Frank, "Discovering the signatures of joint attention in child-caregiver interaction," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36, 2014.

[152] C. D. Heath, H. Venkateswara, T. McDaniel, and S. Panchanathan, "Detecting attention in pivotal response treatment video probes," in *International Conference on Smart Multimedia*, Springer, 2018, pp. 248–259.

[153] K. Campbell, K. L. Carpenter, J. Hashemi, S. Espinosa, S. Marsan, J. S. Borg, Z. Chang, Q. Qiu, S. Vermeer, E. Adler, M. Tepper, H. L. Egger, J. P. Baker, G. Sapiro, and G. Dawson, "Computer vision analysis captures atypical attention in toddlers with autism," *Autism*, pp. 1–10, 2018.

[154] Y. Mitsuzumi, A. Nakazawa, and T. Nishida, "Deep eye contact detector: Robust eye contact bid detection using convolutional neural network," in *BMVC*, 2017.

[155] A. Nakazawa, Y. Mitsuzumi, Y. Watanabe, R. Kurazume, S. Yoshikawa, and M. Honda, "First-person video analysis for evaluating skill level in the humanitude tender-care technique," *Journal of Intelligent & Robotic Systems*, pp. 1–16, 2019.

[156] P. Müller, M. X. Huang, X. Zhang, and A. Bulling, "Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ACM, 2018, p. 31.

[157] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[158] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *ArXiv preprint arXiv:1512.03385*, 2015.

[159] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*, Safety Harbor, Florida, United States of America: ACM, Mar. 2014.

[160] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[161] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[162] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[163] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, 2009, pp. 248–255.

[164] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th international conference on*, IEEE, 2009, pp. 2106–2113.

[165] M. Carpenter, K. Nagell, M. Tomasello, G. Butterworth, and C. Moore, "Social cognition, joint attention, and communicative competence from 9 to 15 months of age," *Monographs of the society for research in child development*, pp. i–174, 1998.

[166] M. Hirotani, M. Stets, T. Striano, and A. D. Friederici, "Joint attention helps infants learn new words: Event-related potential evidence," *Neuroreport*, vol. 20, no. 6, pp. 600–605, 2009.

[167] A. C. MacPherson and C. Moore, "Attentional control by gaze cues in infancy," in *Gaze-Following*, Psychology Press, 2017, pp. 53–75.

[168] T. Charman, J. Swettenham, S. Baron-Cohen, A. Cox, G. Baird, and A. Drew, "Infants with autism: An investigation of empathy, pretend play, joint attention, and imitation.," *Developmental psychology*, vol. 33, no. 5, p. 781, 1997.

[169] A. Senju and M. H. Johnson, "Atypical eye contact in autism: Models, mechanisms and development," *Neuroscience & Biobehavioral Reviews*, vol. 33, no. 8, pp. 1204–1214, 2009.

[170] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[171] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.